Building a 3D Foundation for Spatial AI

MAY 2026

Andrea Vedaldi

UNIVERSITY OF OXFORD

Meta

Why 3D?

Why 3D?

Physical Intelligence

The ability to solve problems in the physical world

[video credit: EPIC Kitchens]

3D for understanding the world



3D for understanding the world



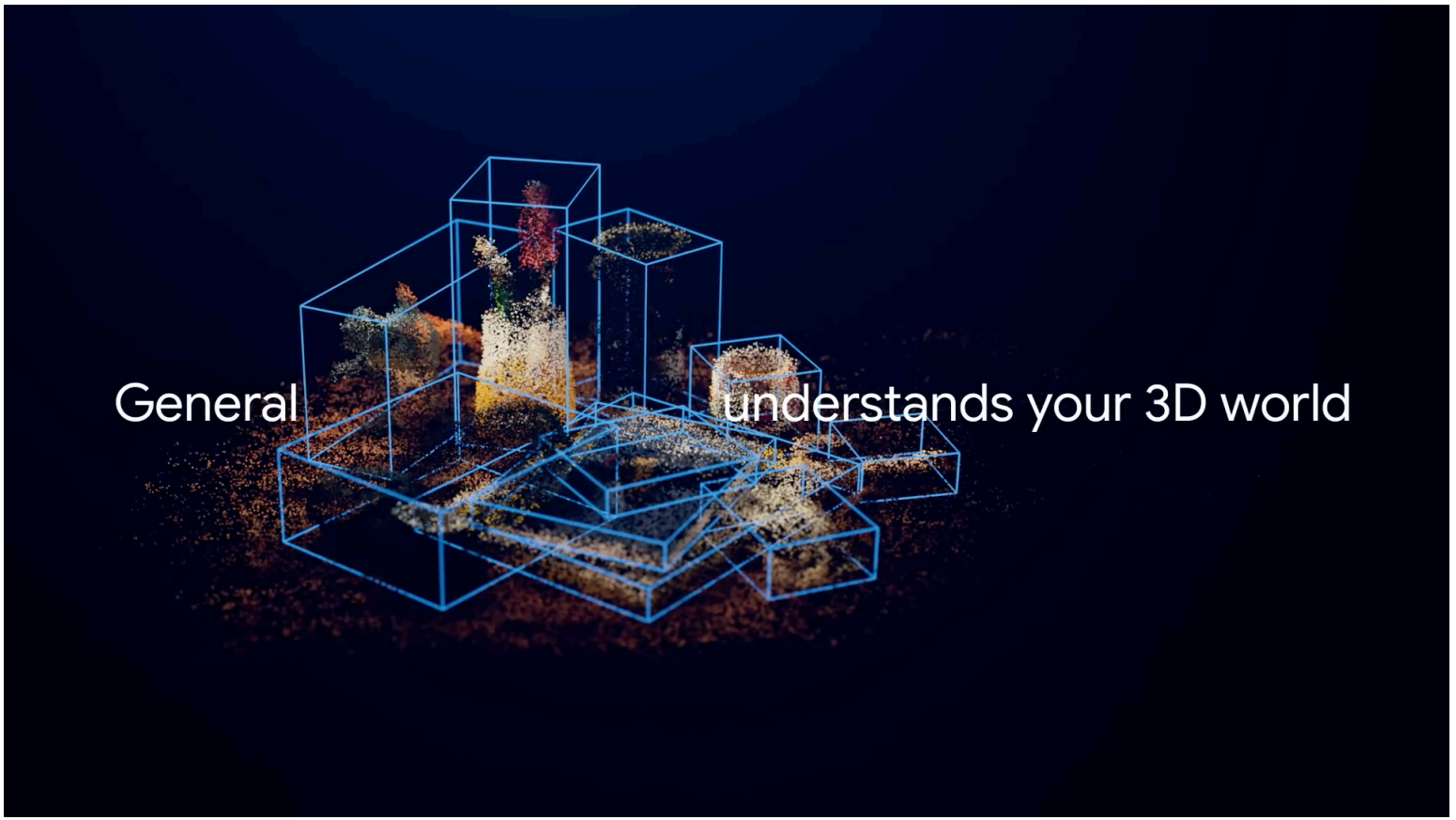
https://epic-kitchens.github.io/epic-fields/

A 3D interpretation **explains away the camera motion**

It also **contextualises** local observations in the wider **world**

EPIC Fields: Marrying 3D geometry and video understanding. Tschernezki, Darkhalil, Zhu, Fouhey, Laina, Larlus, Damen, Vedaldi. NeurIPS, 2023.

3D for acting in the world



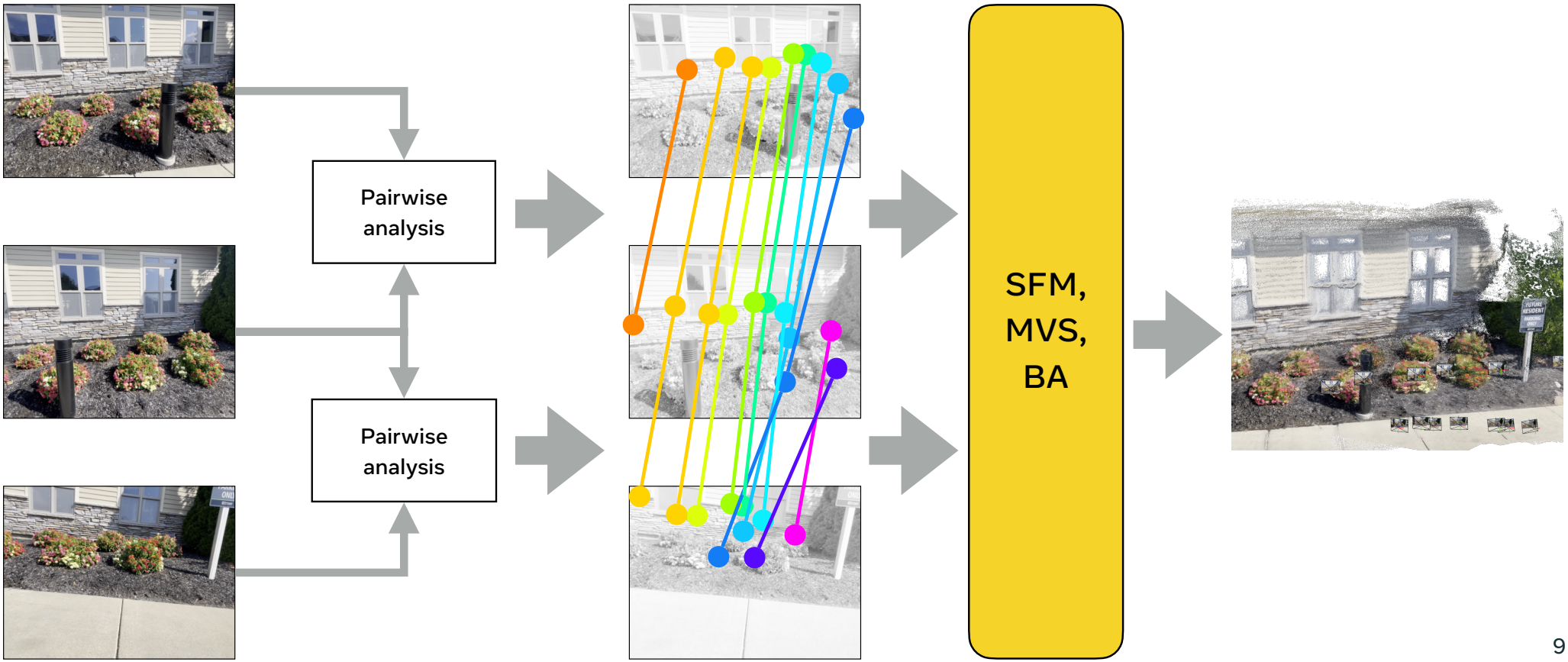
Example: Gemini Robotics

3D for engineering the world



3D transformers

Traditional reconstruction paradigms



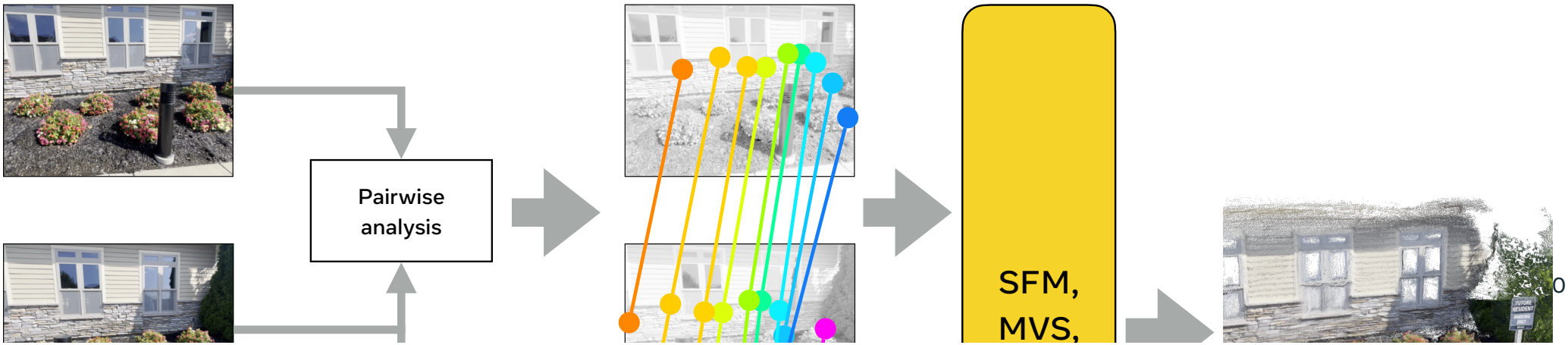
Traditional reconstruction paradigms

Handcrafted or learned
(matchers, depth, pairwise point maps)

SIFT. Lowe. ICCV, 1999.
SuperPoint, DeTone et al., CVPRW 2018
LoFTR. Sun et al., CVPR, 2021
RelPose. Zhang et al. ECCV, 2022
DUSt3R. Wang et al. CVPR, 2024

Optimisation
(visual geometry)

Multiple View Geometry. Hartley&Zisserman, 2000
Photo tourism. Snavely et al. TOG,
COLMAP. Schönberger et al. ECCV, 2016



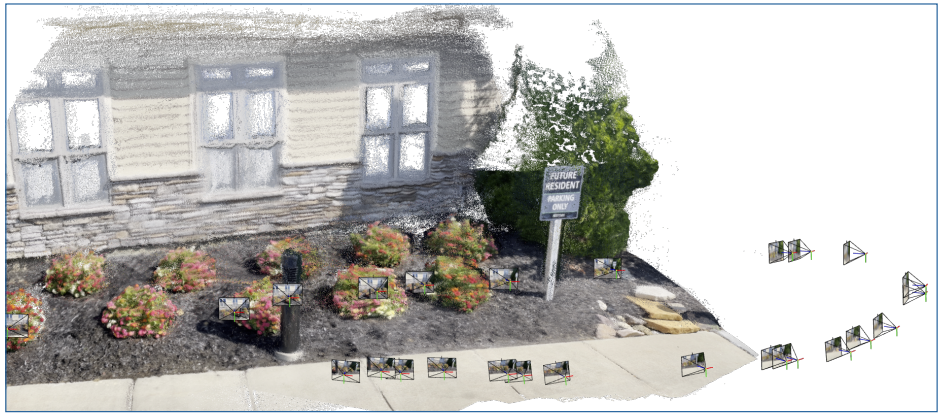
3D transformers

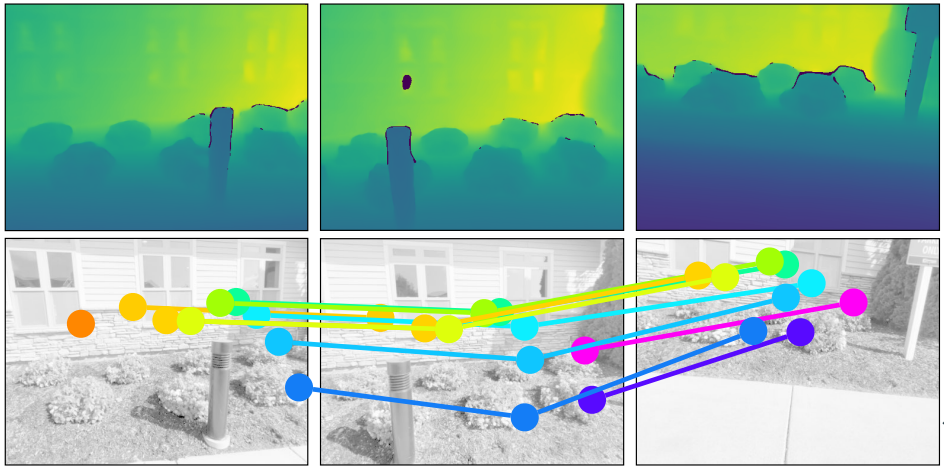




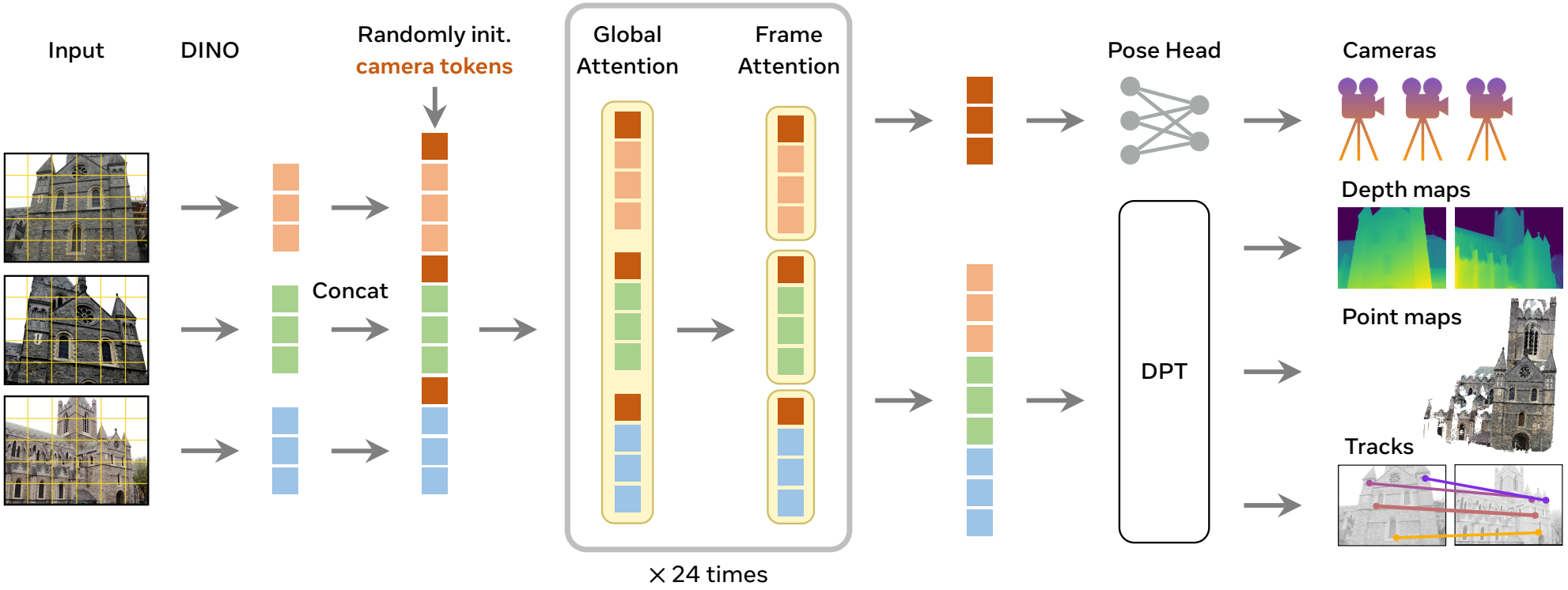


Transformer



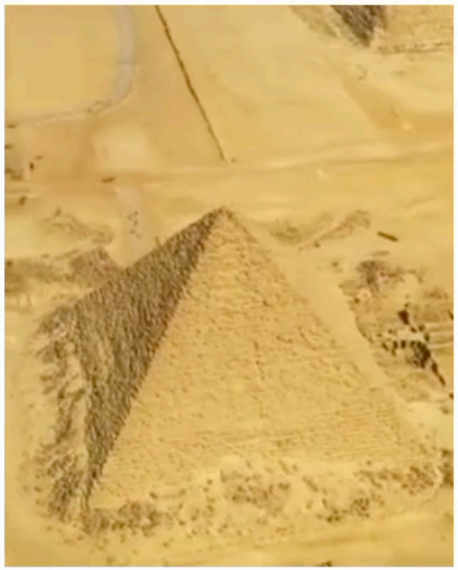


VGGT



VGGT: Visual geometry grounded transformer. Wang, Chen, Karaev, Vedaldi, Rupprecht, Novotny. CVPR, 2025

32 Views



0.51 s



Learnings from VGGT

Transformers can do 3D as well as SfM/BA (in many cases)

3D transformers are good at:

- Multi-view 3D reconstruction
- Monocular depth
- 2D matching & tracking

What else can they do?

What else can 3D transformers do?

New-view synthesis









NVS

Target camera

New View



New-view synthesis



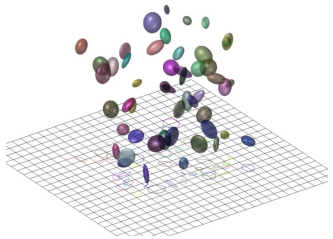






SfM, BA,
NeRF, 3DGS

Explicit 3D
representation



Target camera



Renderer

New View



Let’s use a transformer instead

Poster #112 S3









Encoder (transformer)

Tokens

Target camera

Decoder (transformer)

New View

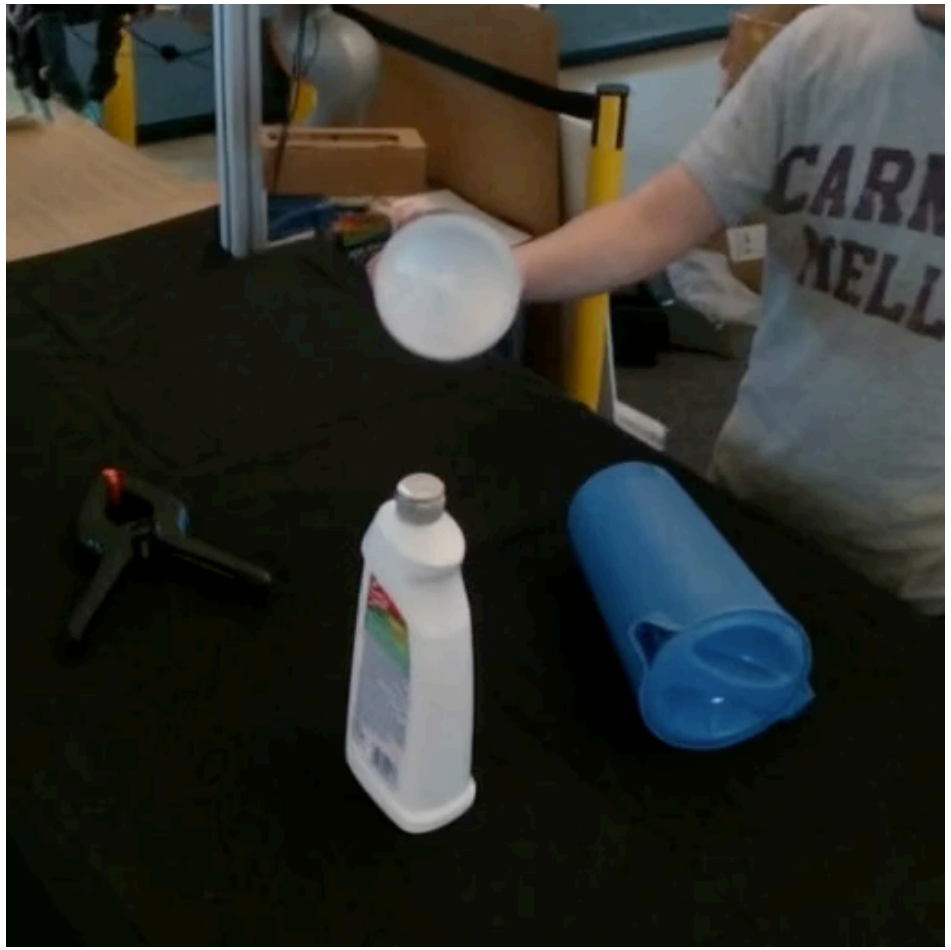


LagerNVS: Latent geometry for fully neural real-time novel view synthesis.
Szymanowicz, Chen, Wang, Rupprecht, Vedaldi. CVPR, 2026


You just proved that these signs work
800-661-6088
CreativeOutdoor.com


You just proved that these signs work
800-661-6088
CreativeOutdoor.com

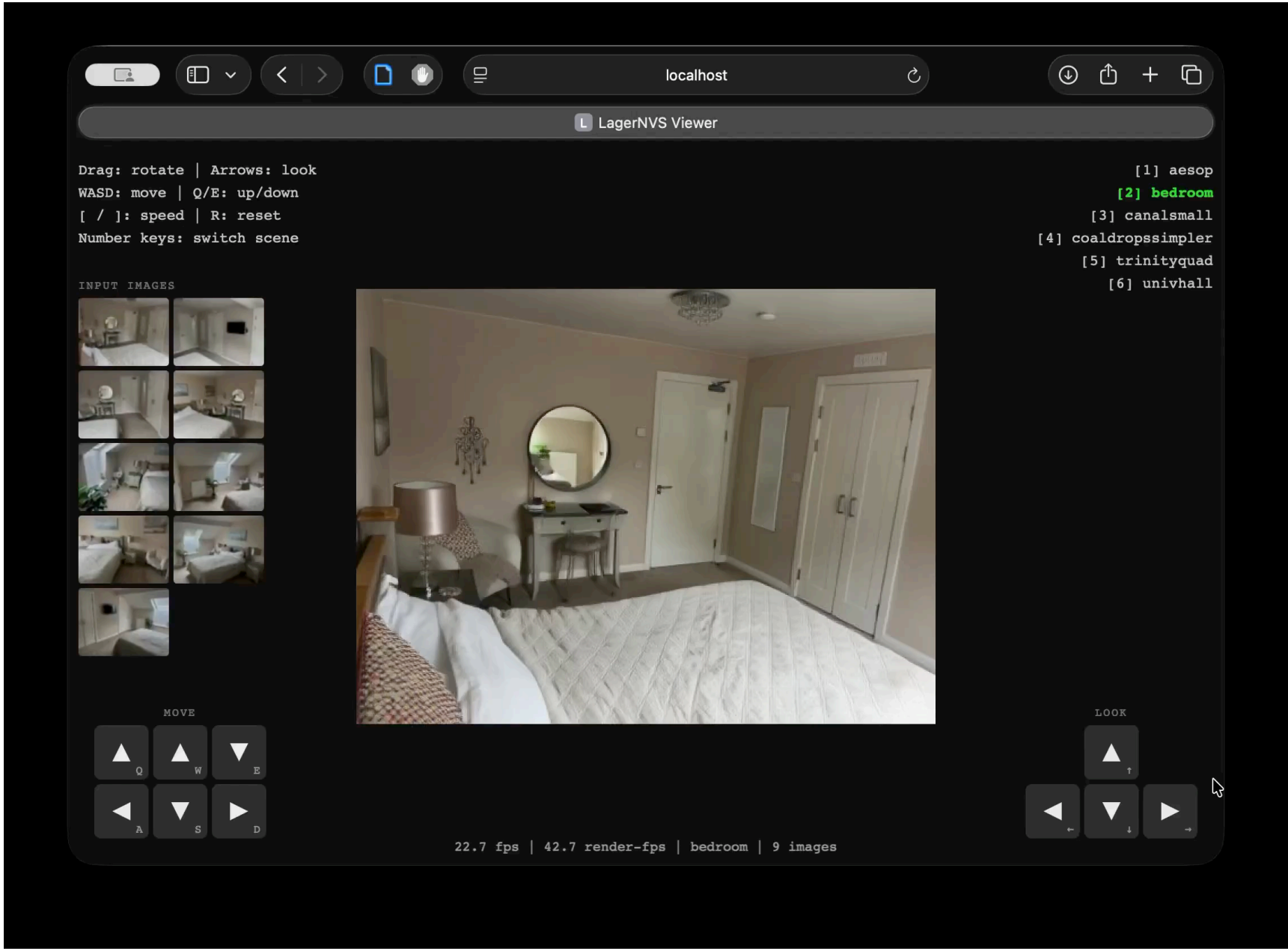




LagerNVS: Latent geometry for fully neural real-time novel view synthesis. Szymanowicz, Chen, Wang, Rupprecht, Vedaldi. CVPR, 2026

Demo

Real time rendering (on a RTX 6000 Blackwell Pro)



3D pre-training









VGGT
(fine-tuned)

3D-aware
tokens

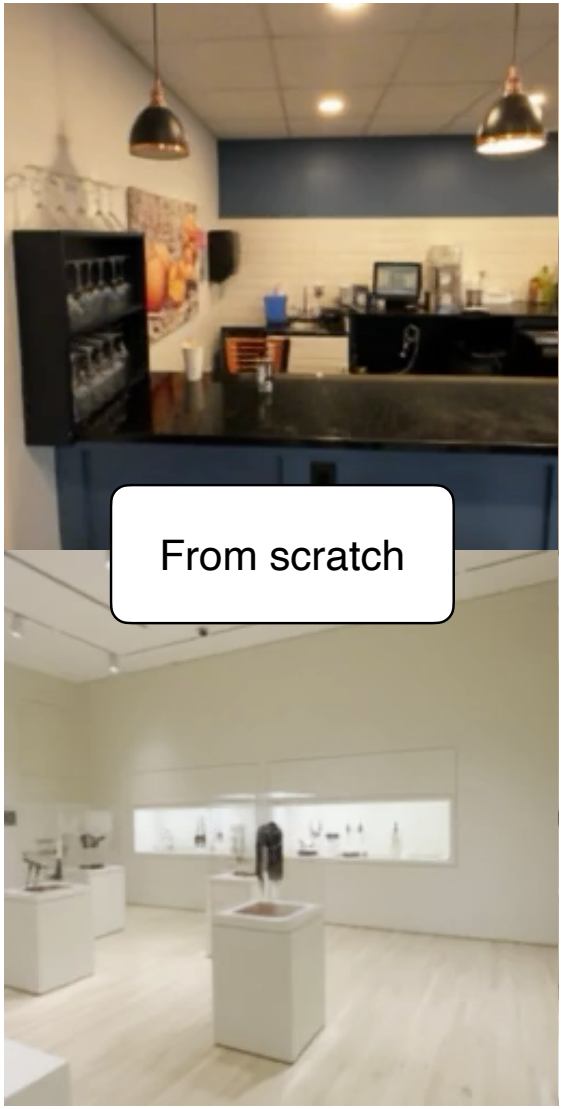
Target camera

Decoder
(transformer)



3D pre-training

Massive impact on performance



LagerNVS: Latent geometry for fully neural real-time novel view synthesis.
Szymanowicz, Chen, Wang, Rupprecht, Vedaldi. CVPR, 2026

Scaling
3D transformers

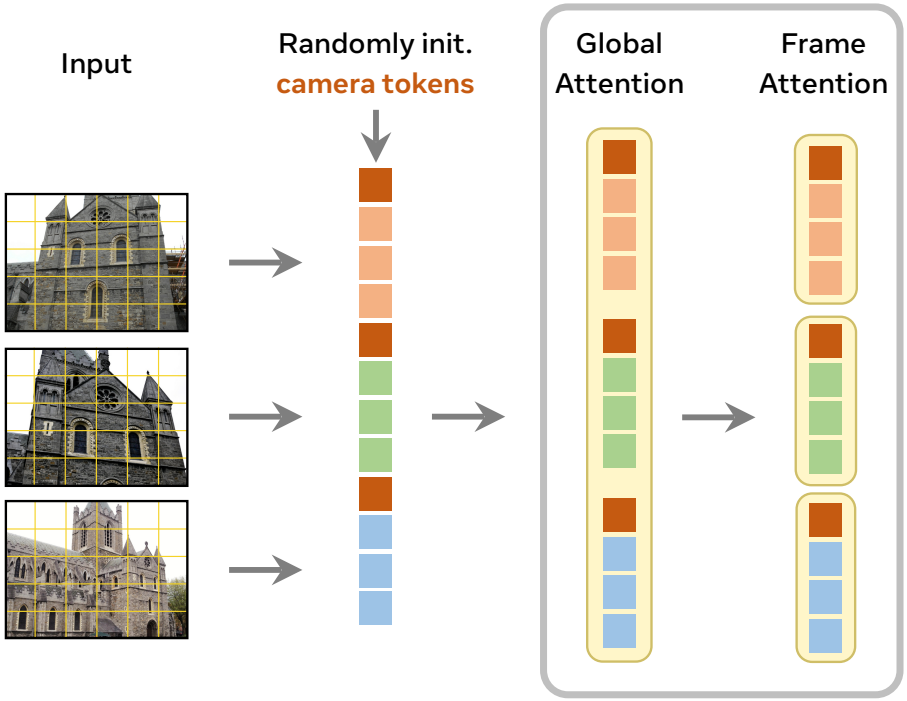
Scaling VGGT

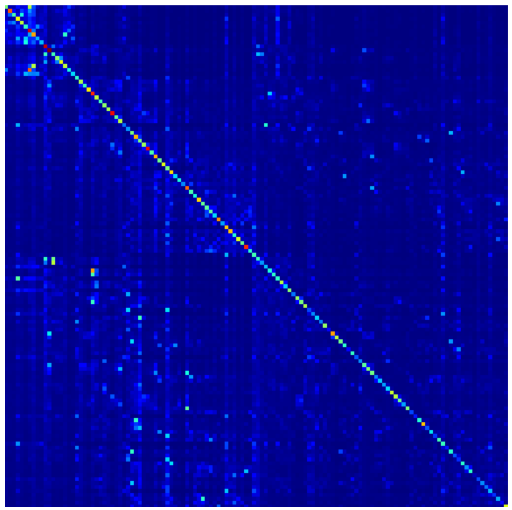
Architecture

Data

VGGT-Ω: Wang, Chen, Zhang, Karaev, Schönberger, Labatut, Bojanowski, Novotny, Vedaldi, Rupprecht. CVPR, 2026

Global attention

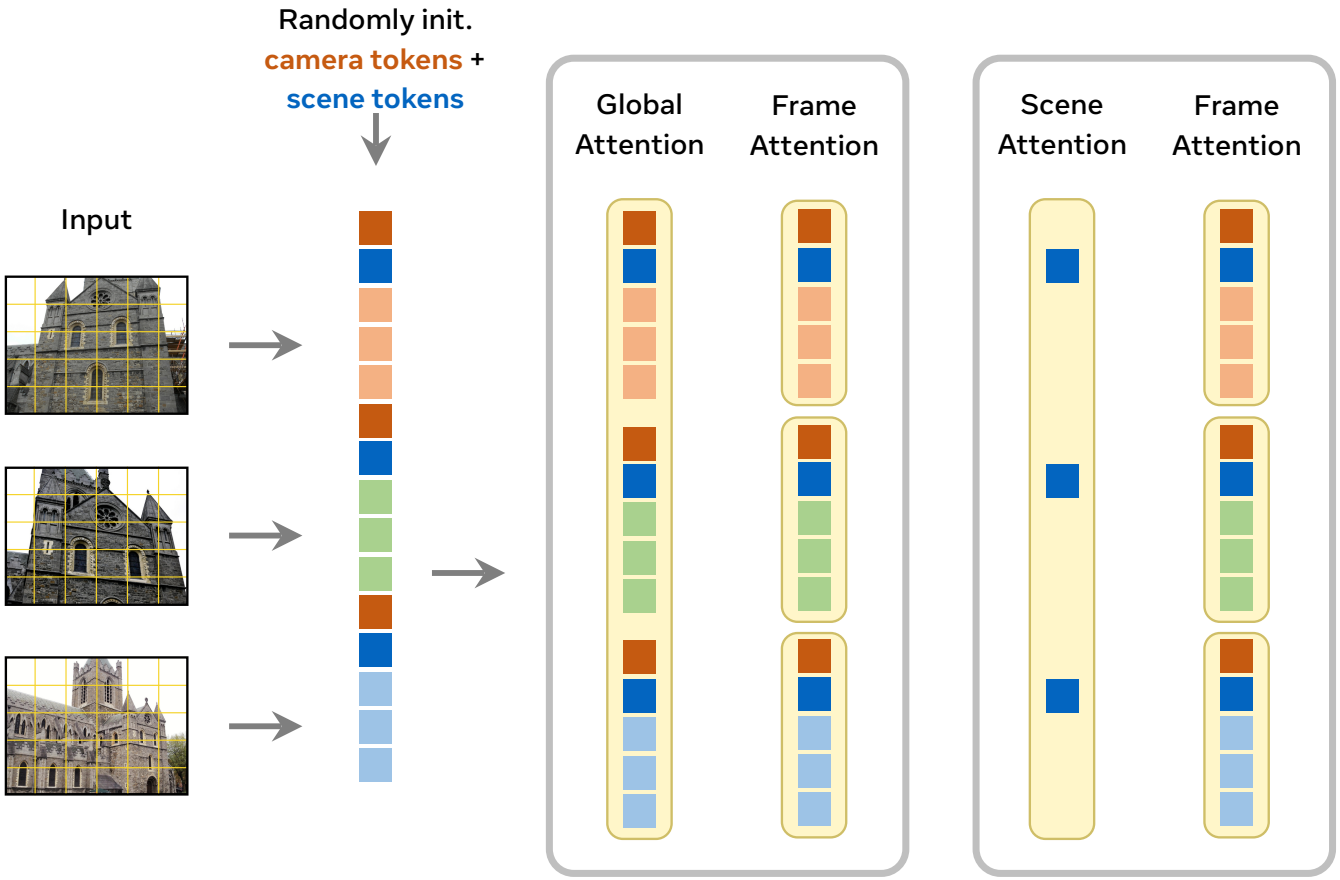




Global attention is a bottleneck

But it is also very sparse…

Scene tokens

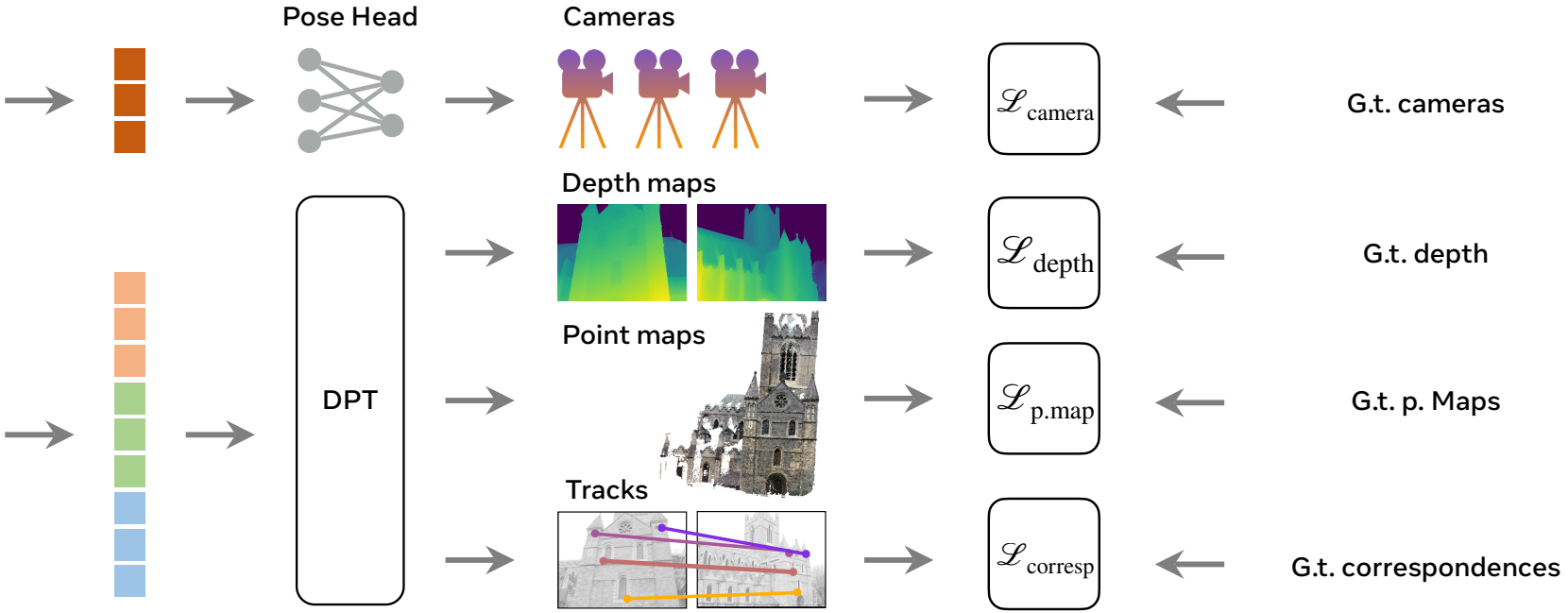


Scene tokens (registers) summarise global information

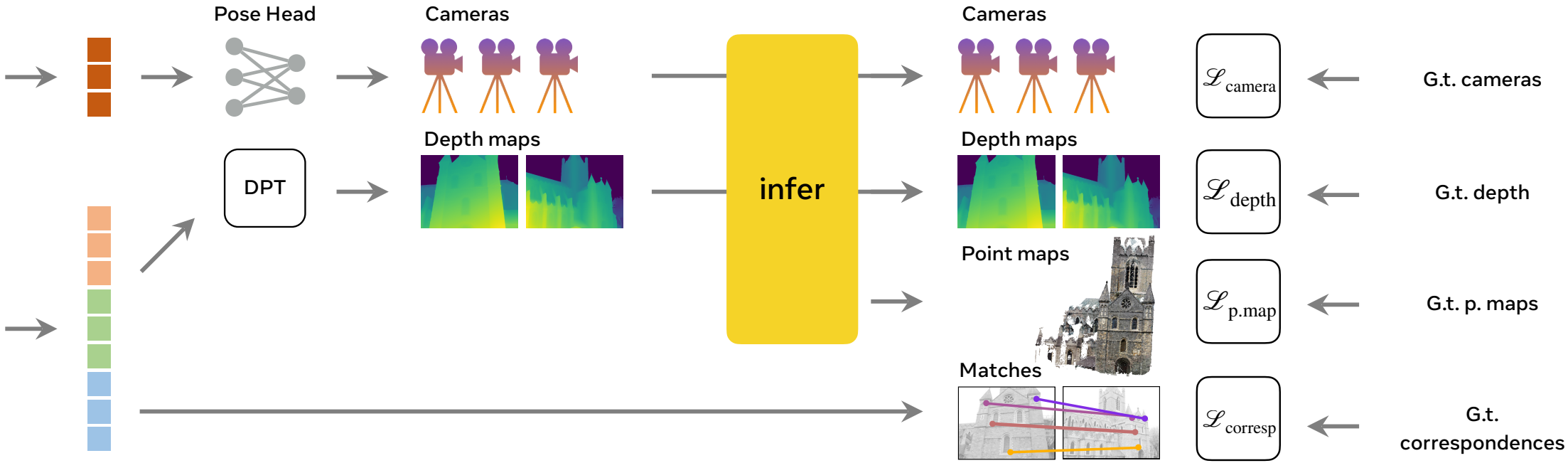
25% of global attention laters are restricted to registers

Saves 13-20% of memory, 20-24% FLOPs, same accuracy

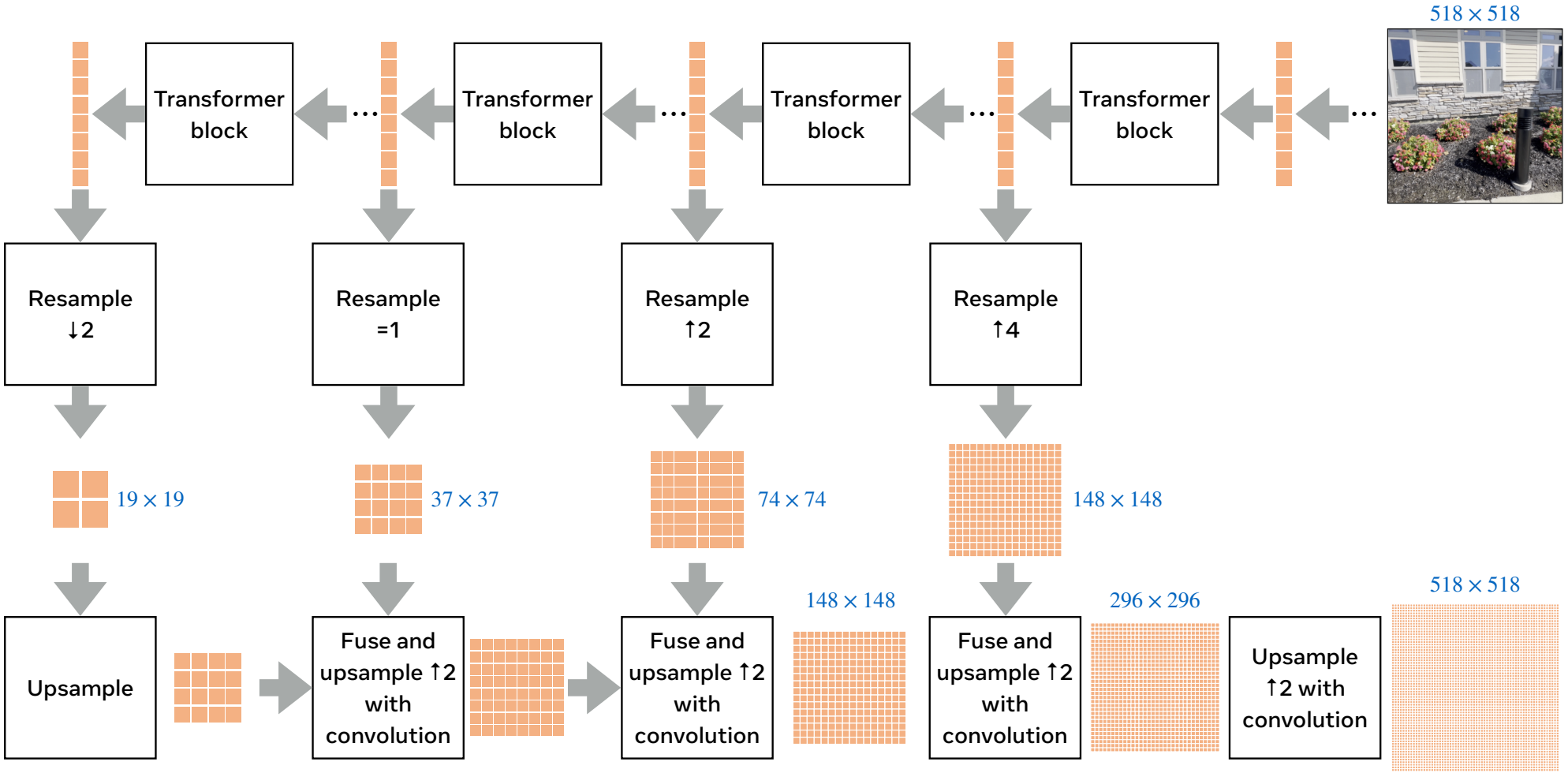
Multi-task supervision



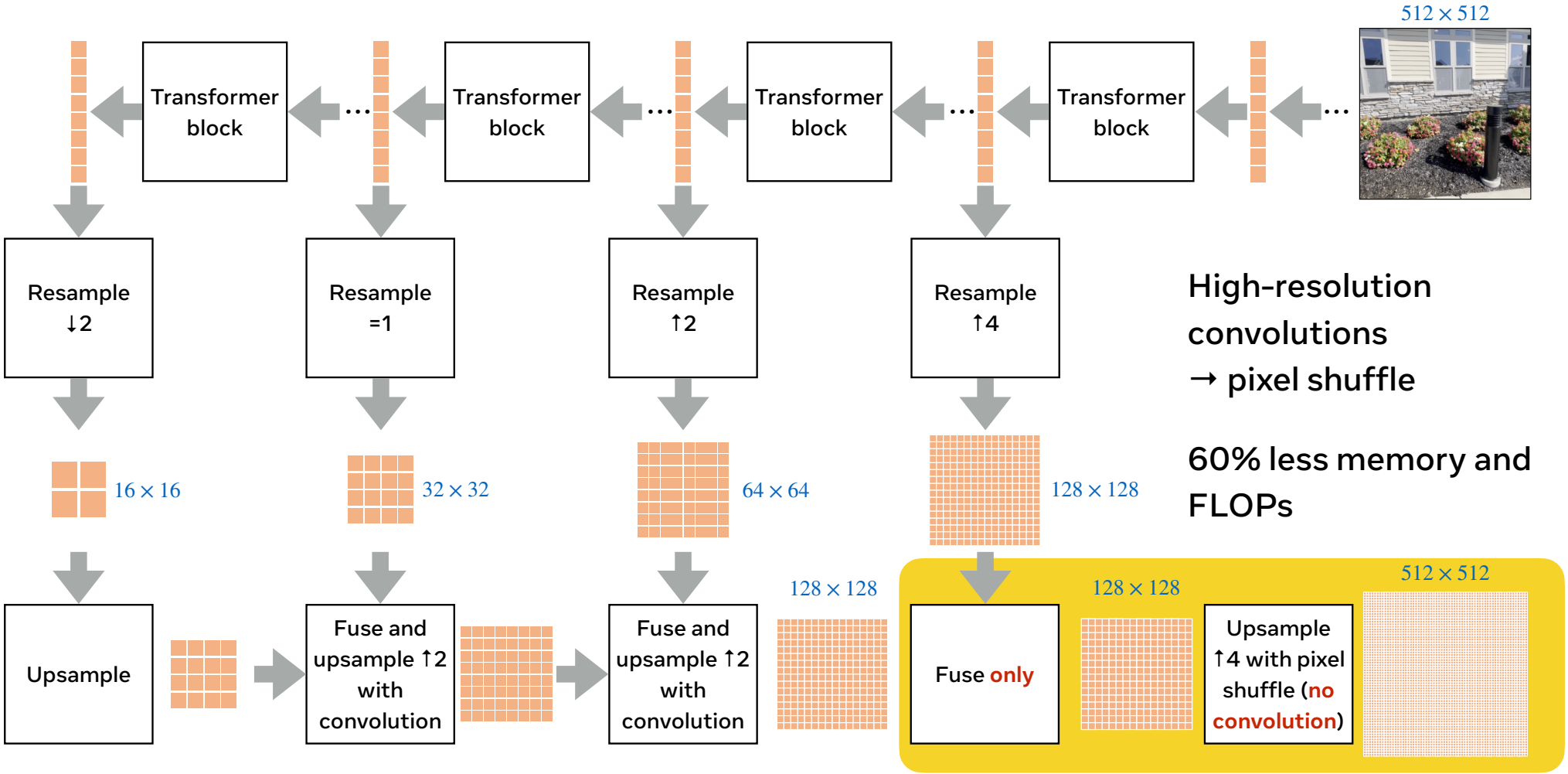
Multi-task supervision with one dense head



DPT decoding



DPT decoding with less convolution



High-resolution convolutions → pixel shuffle

60% less memory and FLOPs

Scaling VGGT

Architecture

Efficiency:

- Scene attention
- Single dense head
- Fewer convolutions

70% memory reduction

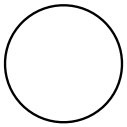
Data

VGGT-Ω: Wang, Chen, Zhang, Karaev, Schönberger, Labatut, Bojanowski, Novotny, Vedaldi, Rupprecht. CVPR, 2026

Scaling data

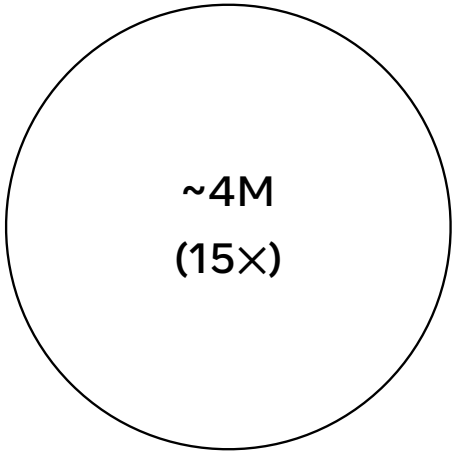
VGGT
labelled data

250k



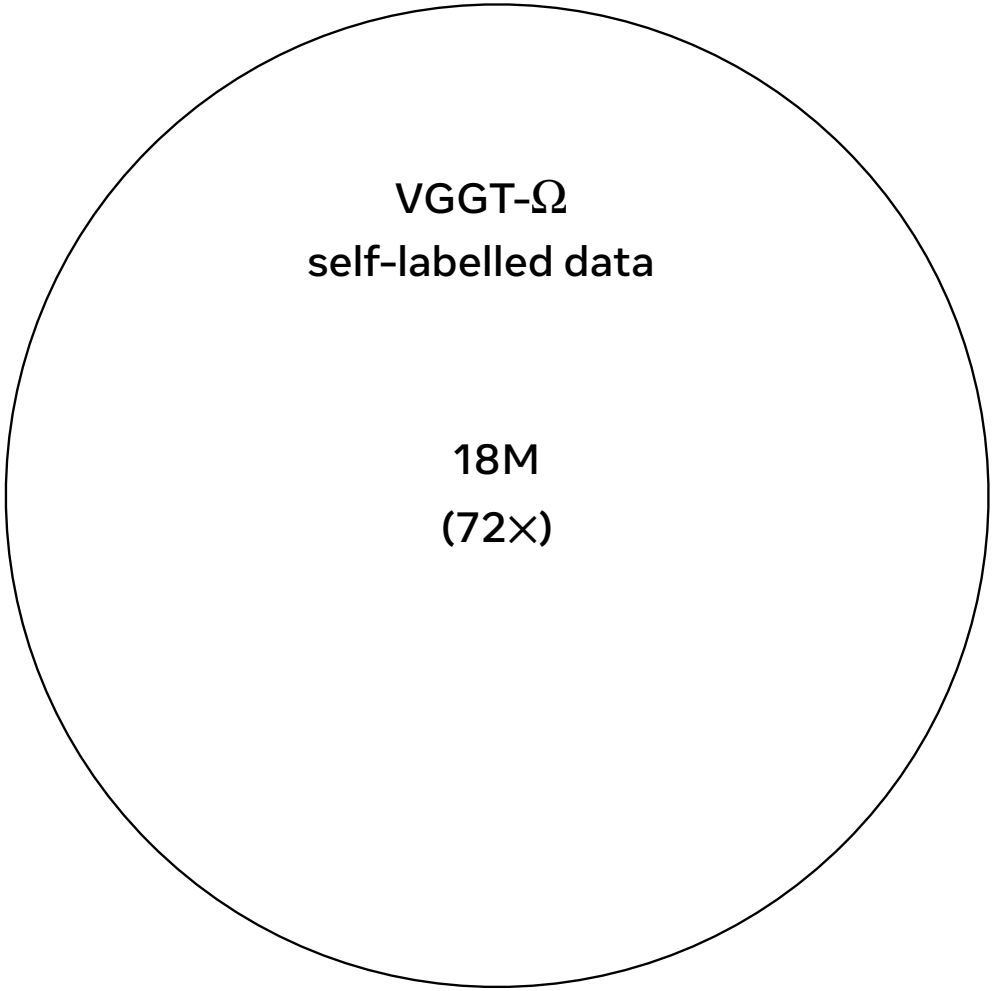
VGGT-Ω
labelled data

~4M
(15×)



VGGT-Ω
self-labelled data

18M
(72×)



Labelled data

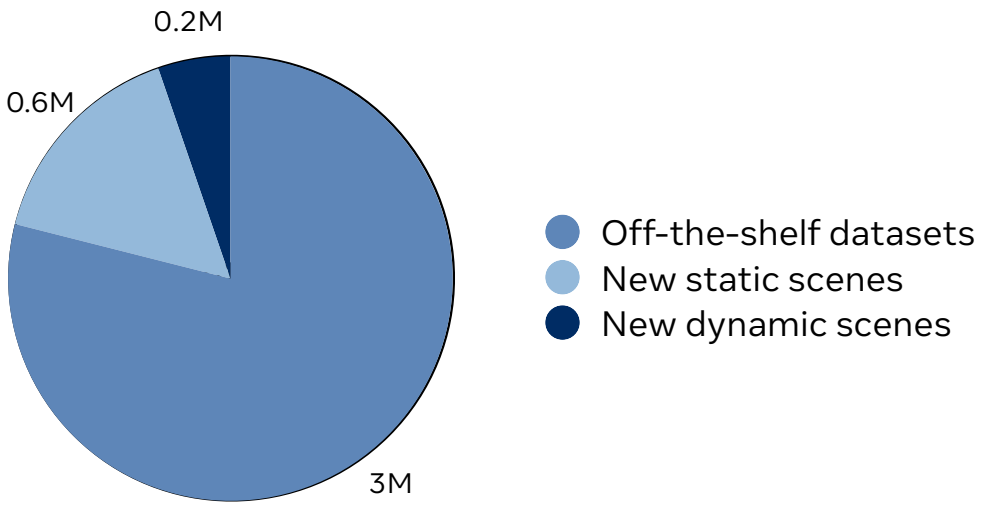
3M videos from 30 datasets

0.8M new videos

- From 40M videos
- Label using COLMAP, VGGT, BA, …
- Strict quality control

Contains dynamic scenes

VGGT-Ω
labelled data



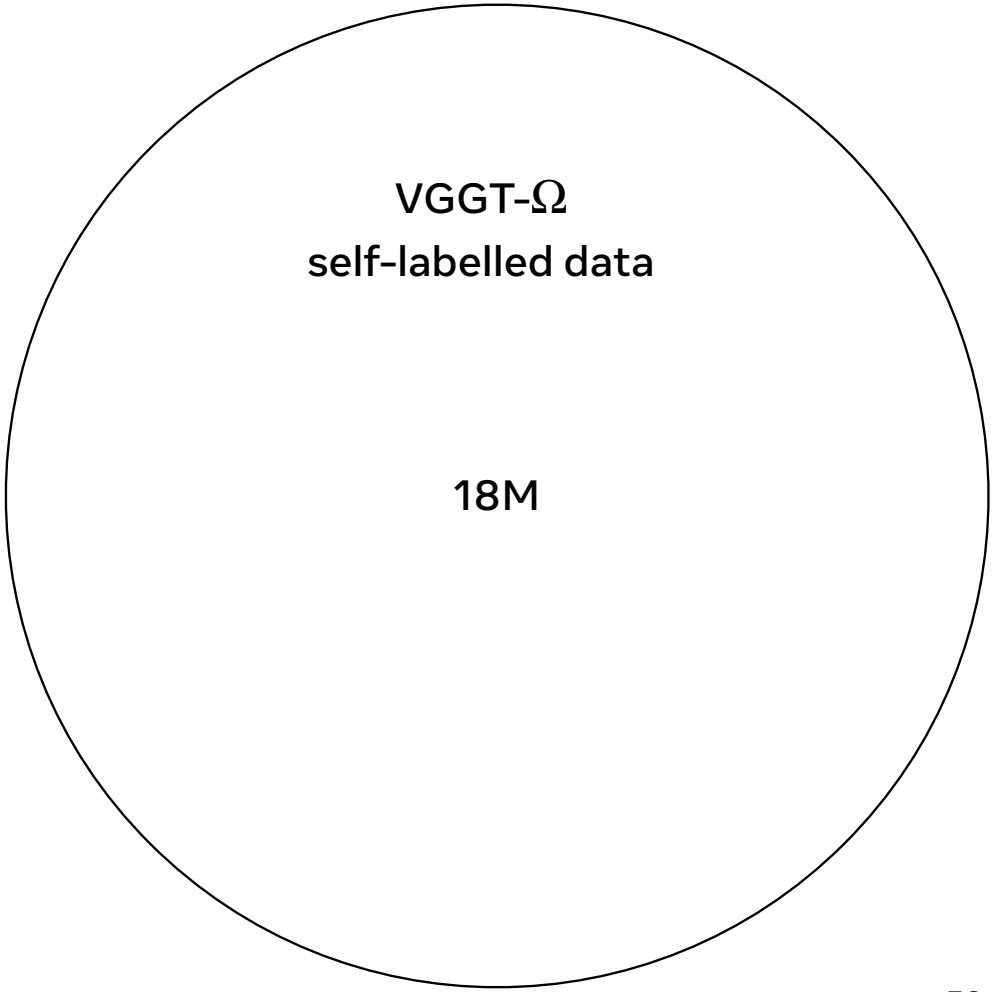
Self-labelled data

18M videos used for self-labelling

Simple protocol inspired by DINO

A student models follows an augmented teacher

Limitation: Mostly useful for generalisation, but does not significantly affect benchmarks



Scaling VGGT

Poster #6 S4

Architecture

Efficiency:

- Scene attention
- Single dense head
- Fewer convolutions

70% overall memory reduction

Data

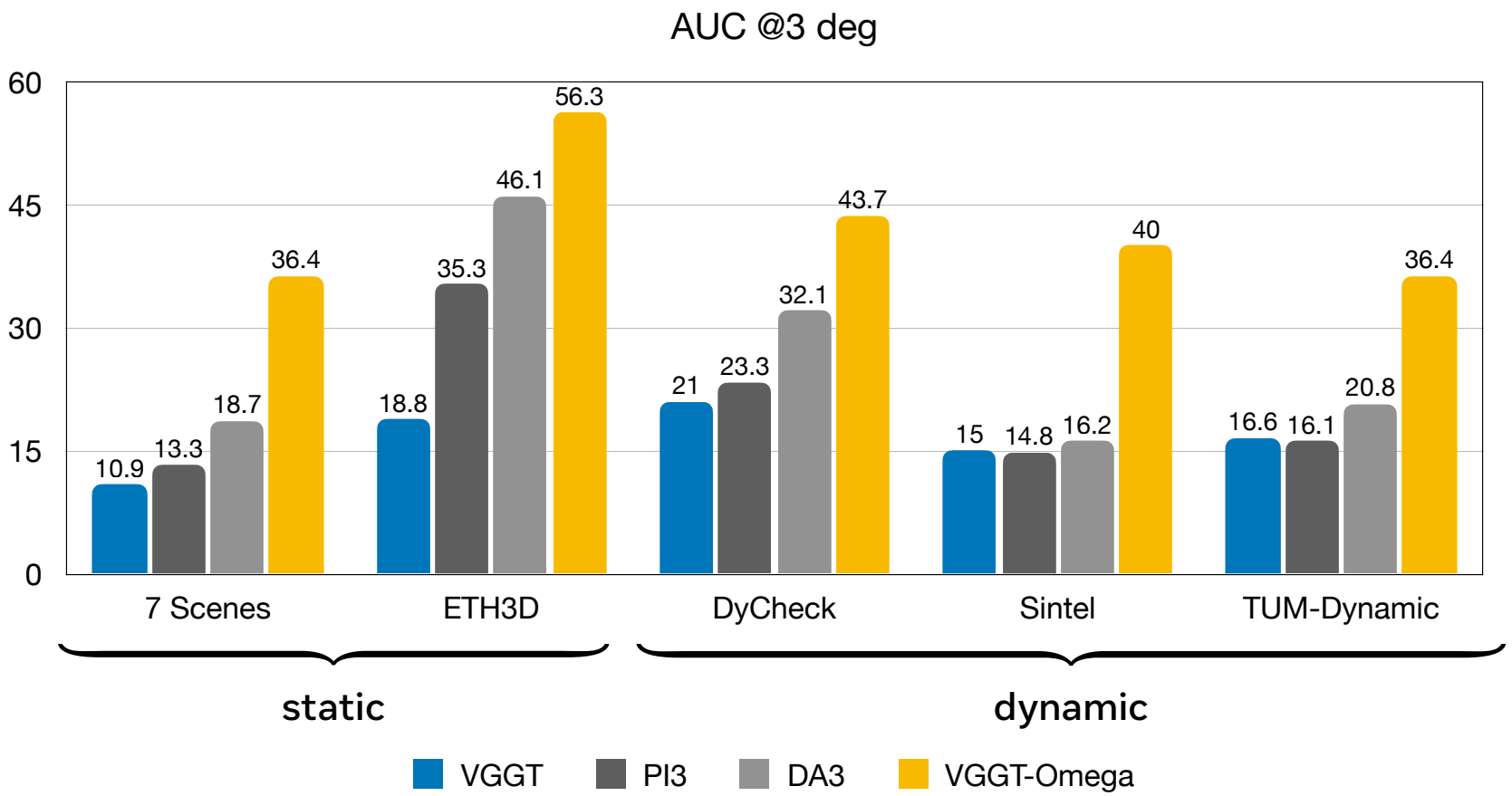
15× more labelled data

(75× new self-labelled data)

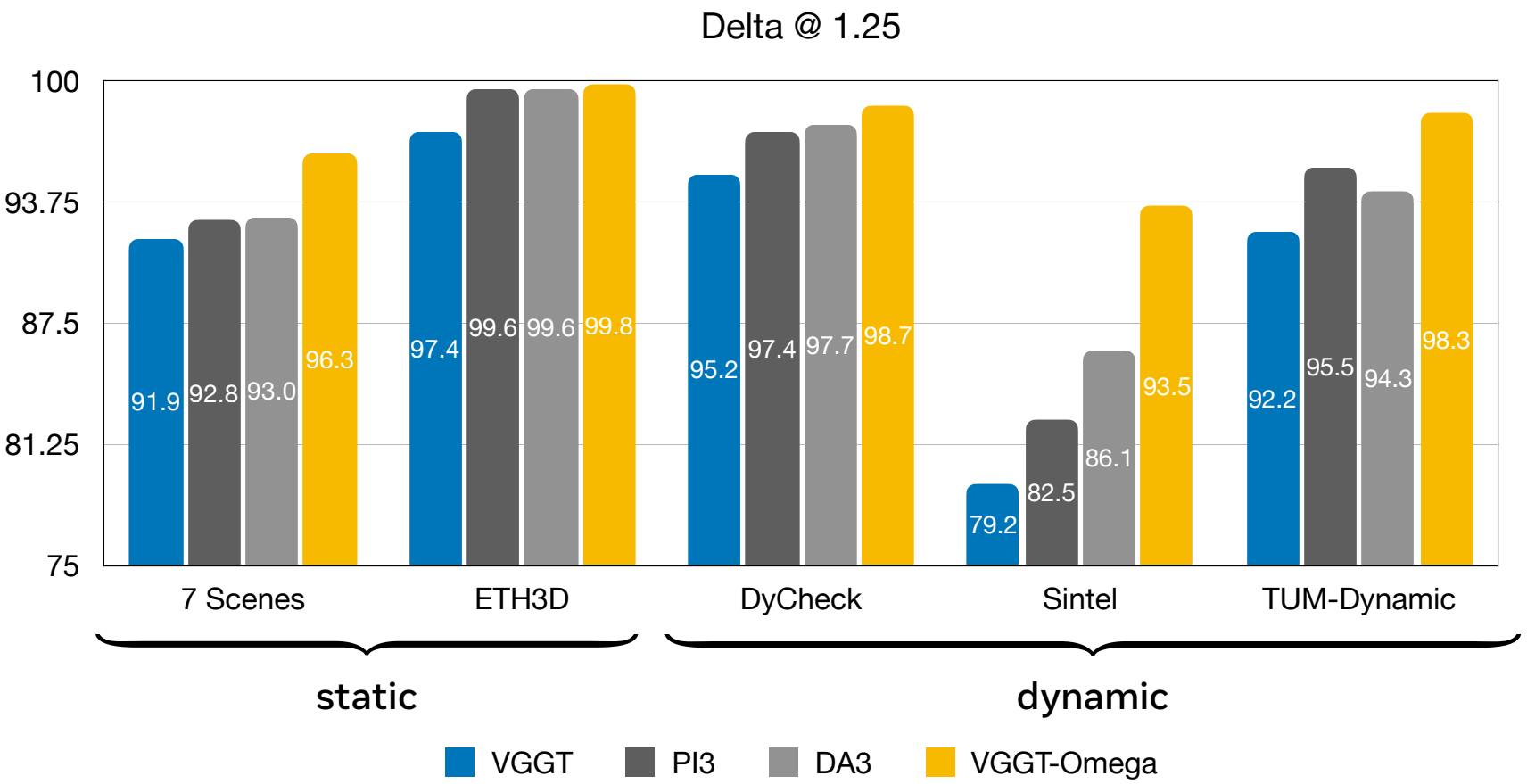
Dynamic scenes

VGGT-Ω: Wang, Chen, Zhang, Karaev, Schönberger, Labatut, Bojanowski, Novotny, Vedaldi, Rupprecht. CVPR, 2026

Camera reconstruction



Depth reconstruction



VGGT-Ω



VGGT-Ω



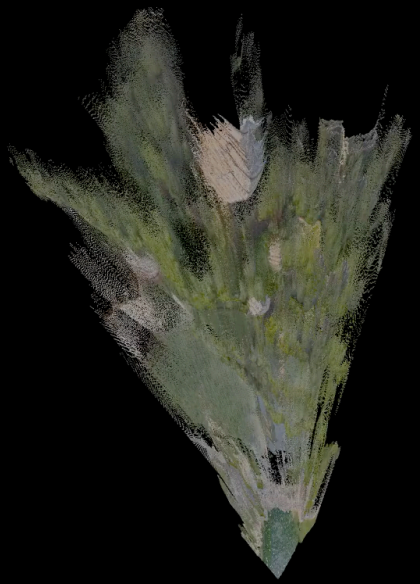


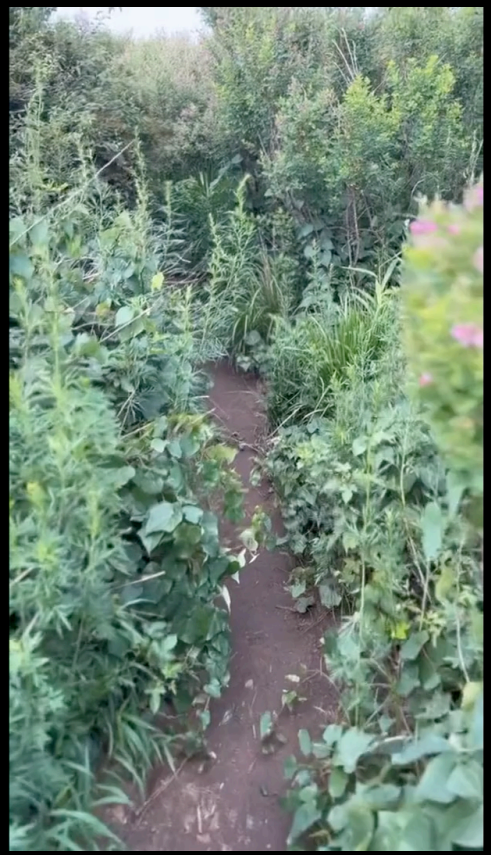


VGGT-Ω

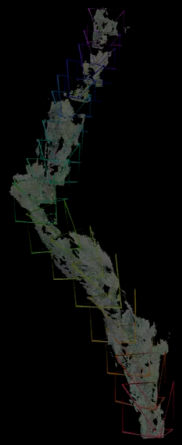


Depth Anything 3

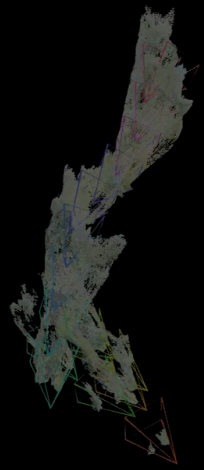




VGGT-Ω



Depth Anything 3





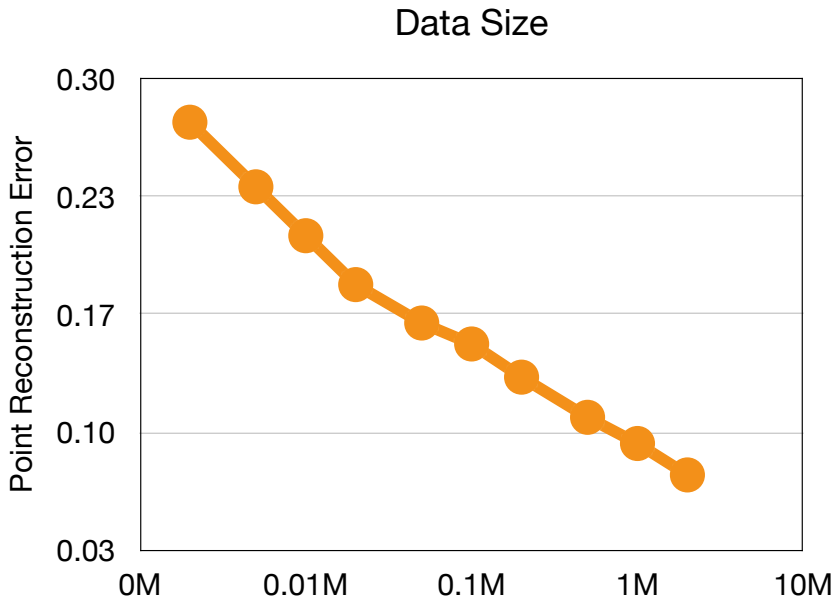
VGGT-Ω

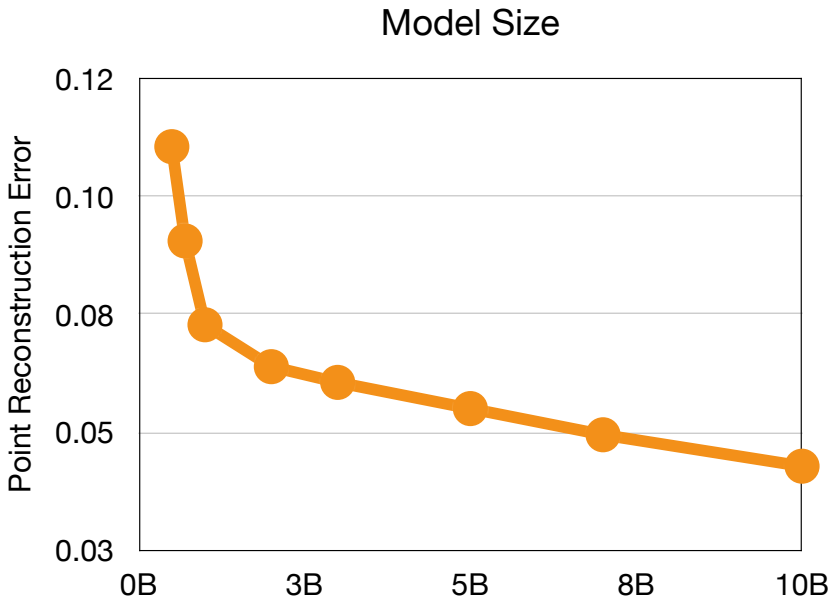


Depth Anything 3



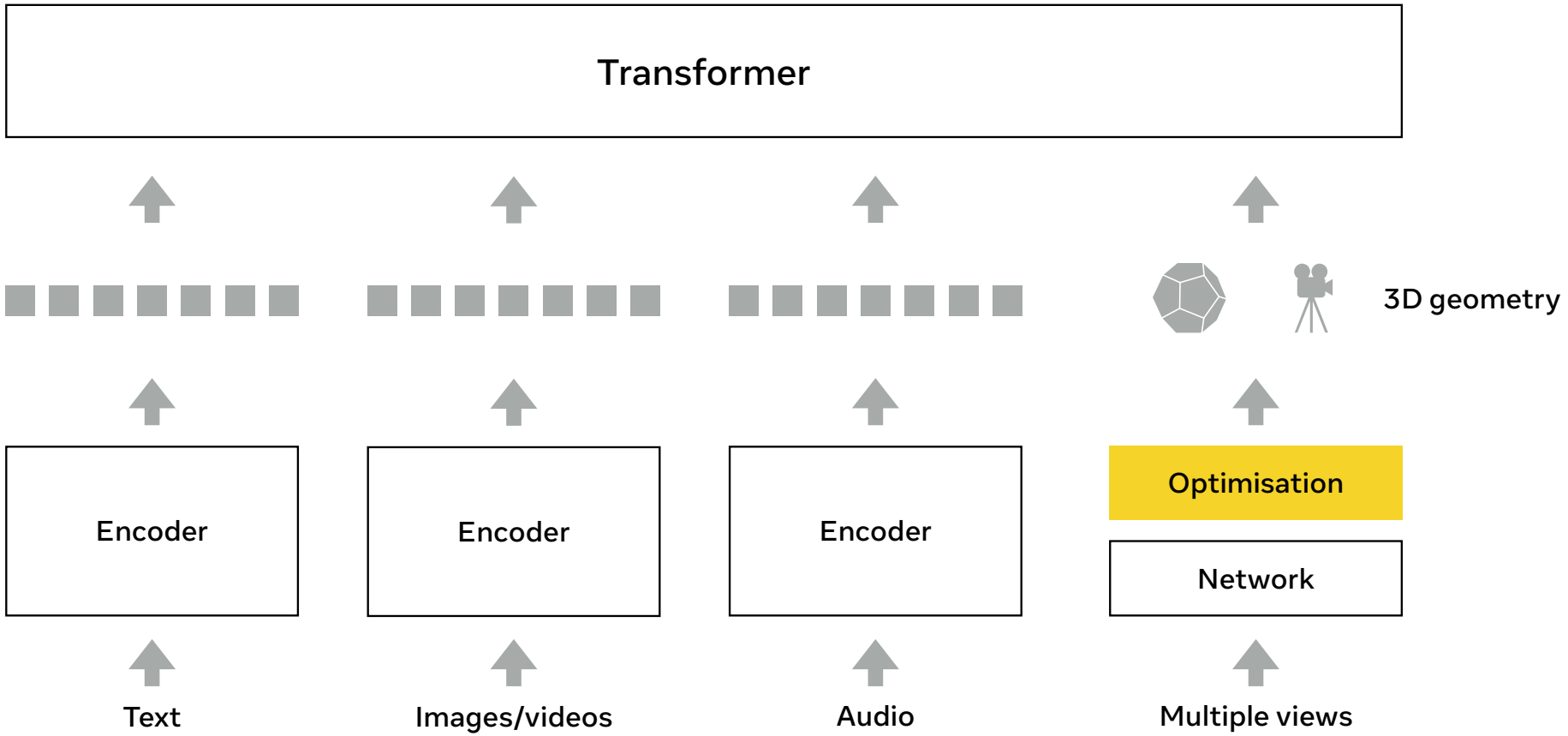
Scaling behaviour



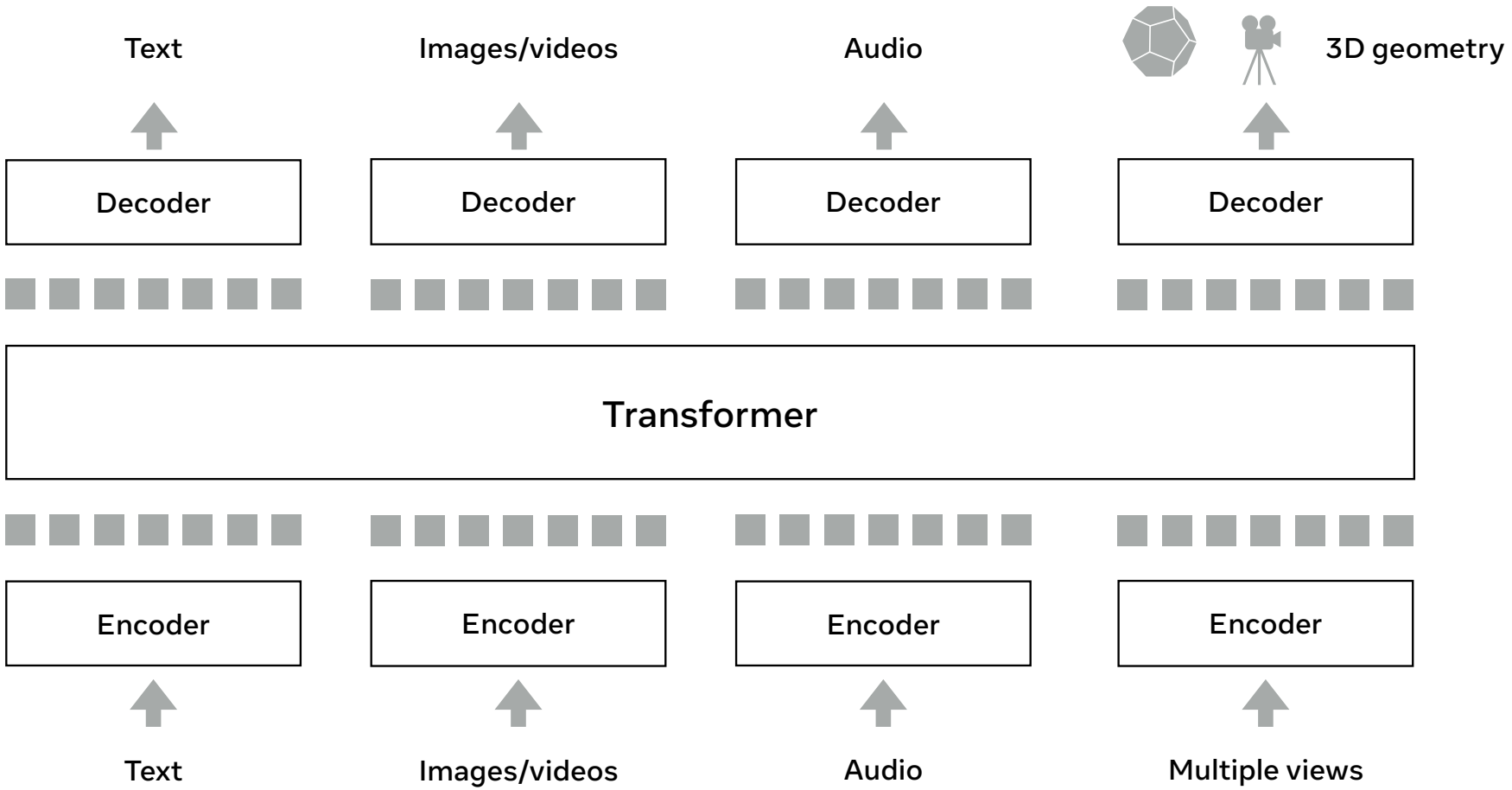


3D transformers as a foundation

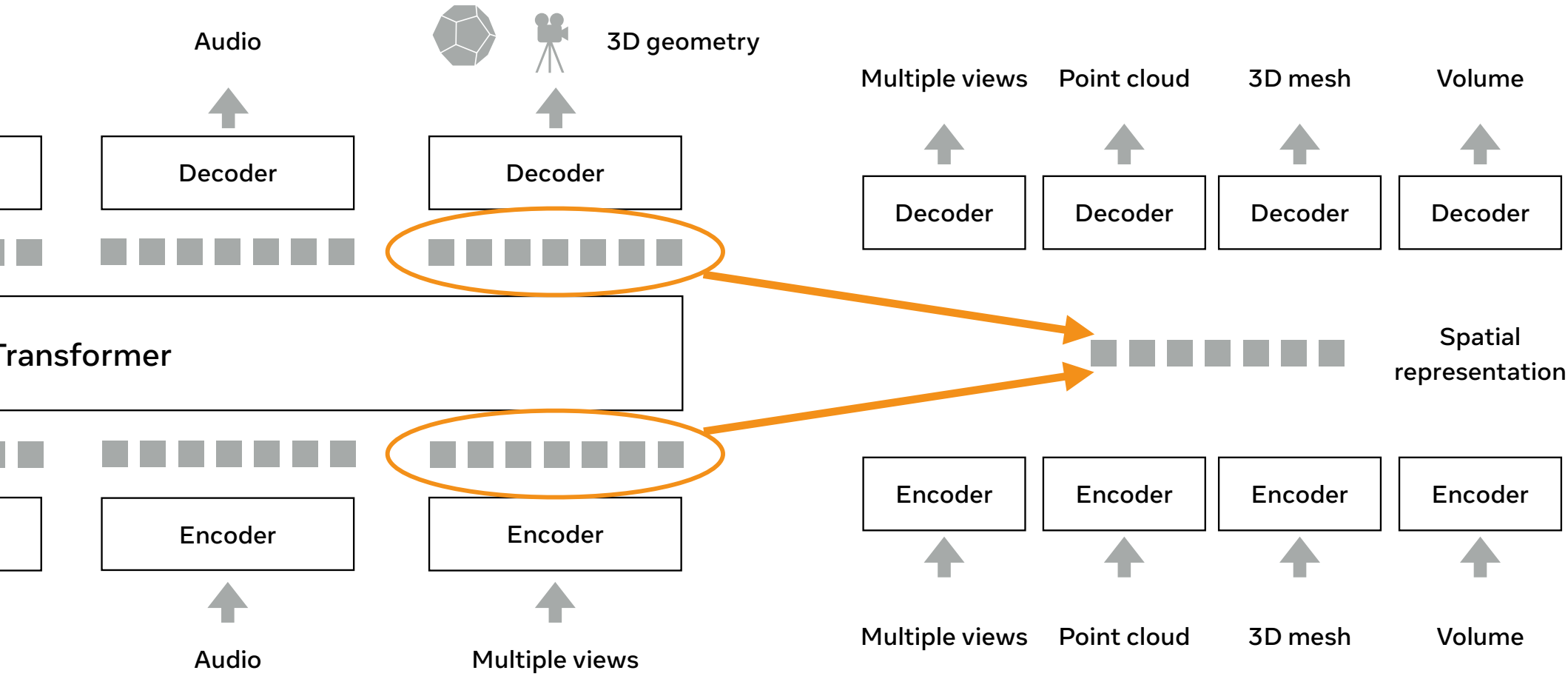
3D in AI


Transformer
3D geometry
Encoder
Encoder
Encoder
Optimisation
Network
Text
Images/videos
Audio
Multiple views

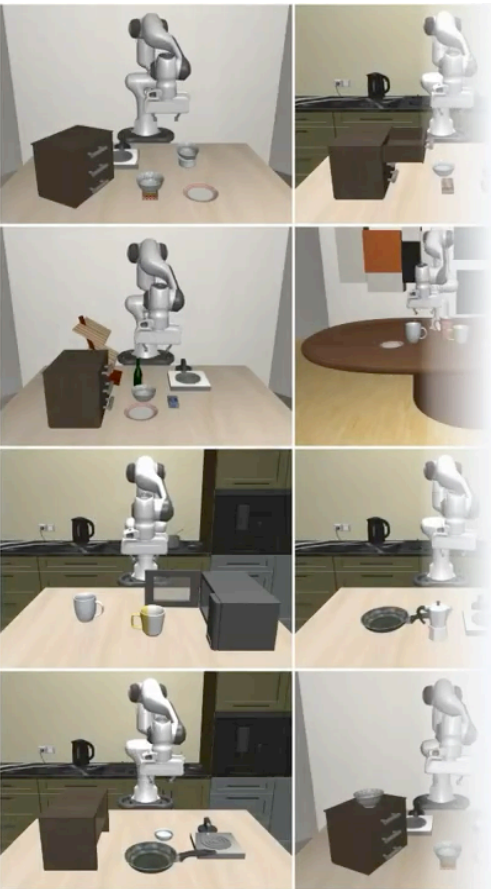
A better way


Text
Images/videos
Audio
3D geometry
Decoder
Decoder
Decoder
Decoder
Transformer
Encoder
Encoder
Encoder
Encoder
Text
Images/videos
Audio
Multiple views

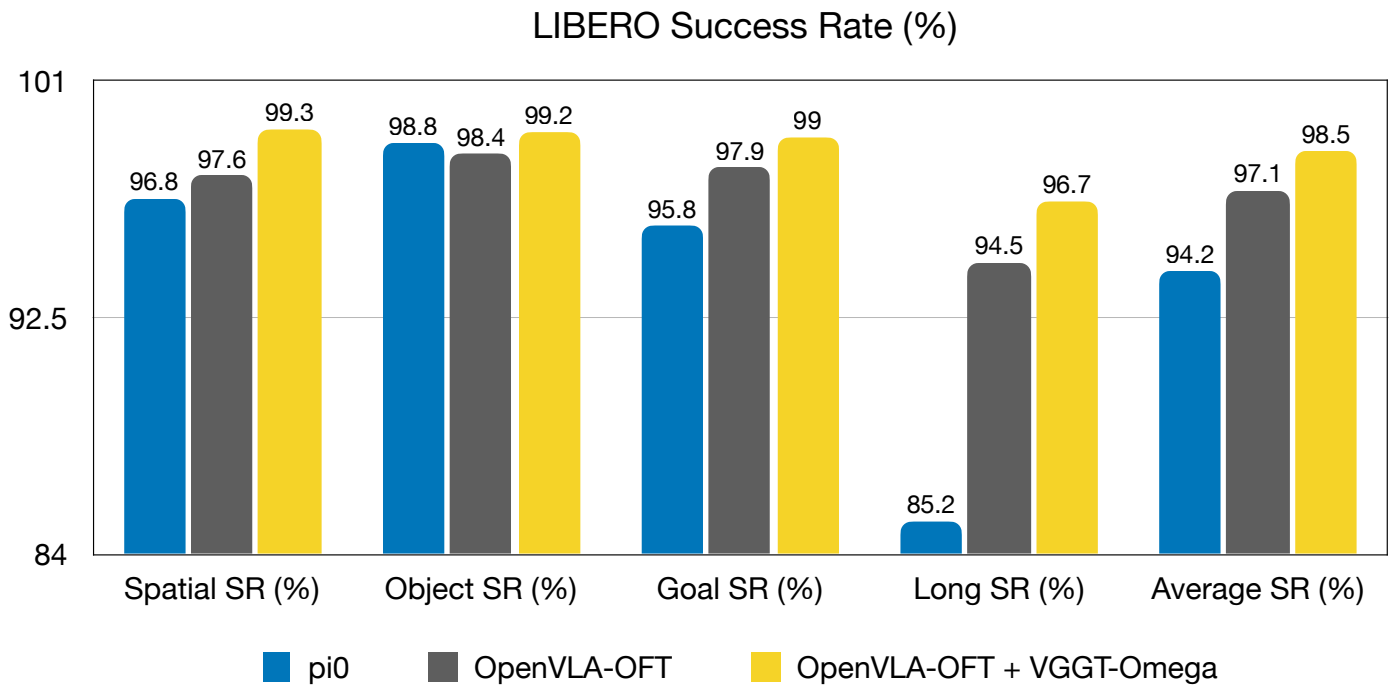
A better way


Audio
3D geometry
Multiple views
Point cloud
3D mesh
Volume
Decoder
Decoder
Decoder
Decoder
Decoder
Decoder
Transformer
Spatial representation
Encoder
Encoder
Encoder
Encoder
Encoder
Encoder
Audio
Multiple views
Multiple views
Point cloud
3D mesh
Volume

VGGT-Ω (frozen) + OpenVLA



[LIBERO]

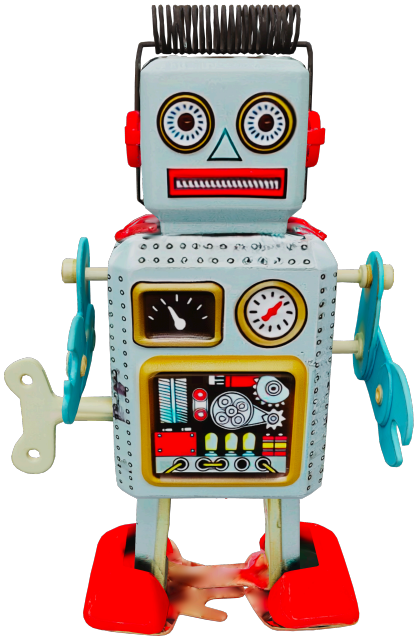


Fine-tuning vision-language-action models... Kim, Finn, Liang. Proc. RSS, 2025


3D reconstruction

Text-to-3D

"A vintage toy robot"



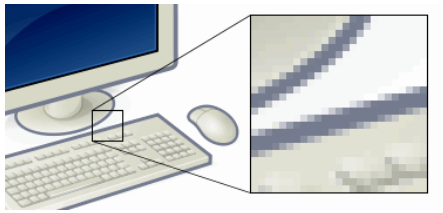
Like image generation, but in 3D

https://developers.meta.com/horizon/blog/worlds/AssetGen2/

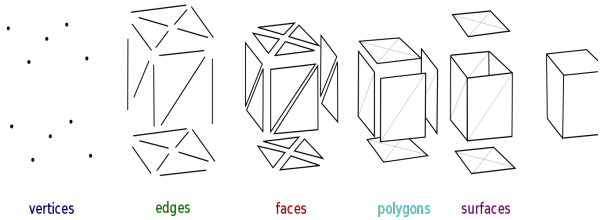
[images from Wikimedia]

Useful 3D generation entails structure

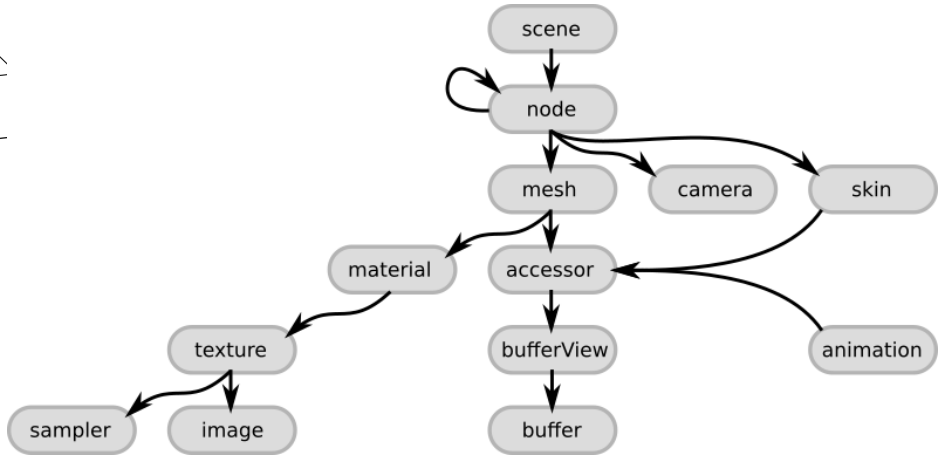
2D pixels



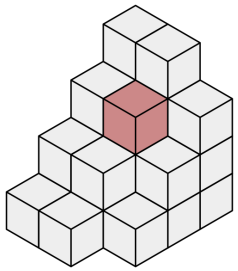
3D cells and topology



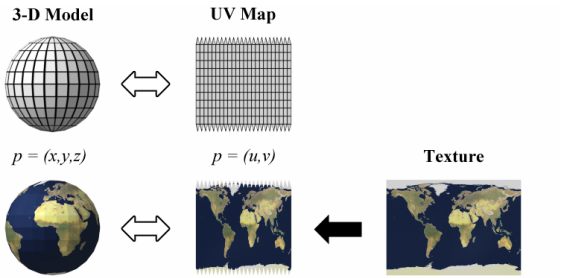
GLTF file structure



3D voxels

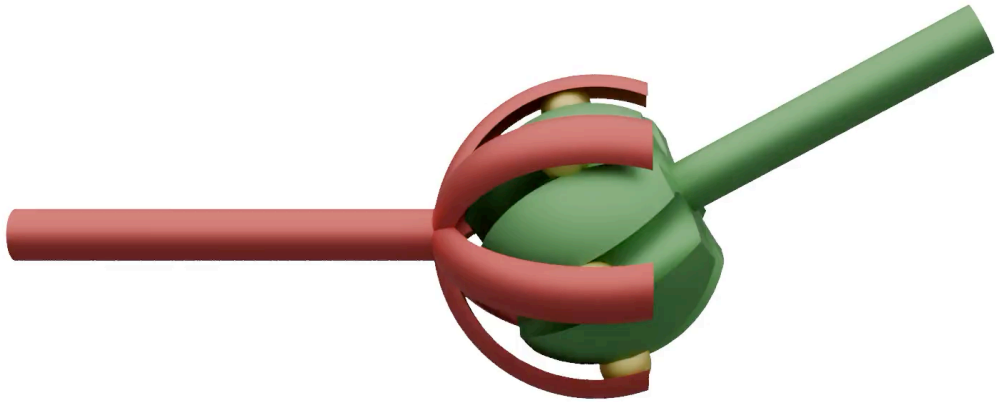


UV maps and atlases



• • •

Objects like mechanisms are poorly modelled as “3D images/ videos”...



...but can be constructed compactly via code

```
rotate([0, 0, 0])
rotate([rot, 0, 0]) {
  outerShell();
}
color([1, 1, 0])
rotate([0, ang / 2, 0])
rotate([rot, 0, 0]) {
  balls();
}
color([0, 1, 0])
rotate([0, ang, 0])
rotate([rot, 0, 0]) {
  gear();
}

module outerShell() {
  for (i = [0:1:n]) {
    difference() {
      rotate([360 / n * i])
      rotate_extrude(convexity=convexity)
      translate([r1, 0, 0])
      circle(r=r3 + outerShellThickness);

      rotate([360 / n * i])
      rotate_extrude(convexity=convexity)
      translate([r1, 0, 0])
      circle(r=r3);

      translate([r2 + r3 + 2
        + outerShellOverhang, 0, 0])
      cube(size=(r2 + 2) * 2,
        center=true);

      sphere(r=r1 + outerShellClearance);
    }
  }

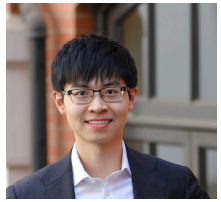
  translate([-r2 - h / 2, 0, 0])
  rotate([90, 90, 90])
  cylinder(h=h, r=r4, center=true);
}

module balls() {
  for (i = [0:1:n * 2]) {
    rotate([360 / n * i / 2])
    translate([0, r1, 0])
    sphere(r=r3);
  }
}

module gear() {
  intersection() {
    difference() {
      rotate([0, 90, 0])
      sphere(r=r1);
      for (i = [0:1:n]) {
      rotate([360 / n * i])
      rotate_extrude(convexity=convexity)
      translate([r1, 0, 0])
      circle(r=r3);
    }
  }
  cube(size=[(r1 - c) * 2, (r1 + 2) * 2,
    (r1 + 2) * 2], center=true);
  }

  translate([h / 2, 0, 0])
  rotate([90, 90, 90])
  cylinder(h=h, r=r4, center=true);
}
```


[Michael Frey / Wikimedia Commons]



Shangzhe Wu
@ Cambridge

Can LLMs construct 3D objects?

LLM → 3D model

"Articulating desk lamp"

LLM

3D model file

```
<robot name="robot">
 <!-- links -->
 <link name="root">
   <inertial>
     <origin xyz="0 0 0.5" rpy="0 0
0"/>
     <mass value="1"/>
     <inertia ixx="100"  ixy="0"
ixz="0" iyy="100" iyz="0"
izz="100" />
   </inertial>
   <visual>
     <origin xyz="0.0 0.0 0.0"
rpy="0.0 0.0 0.0" />
     <geometry>
       <box size="0.2 0.2 0.4" />
     </geometry>
     <material name="Cyan">
       <color rgba="1.0 1.0 1.0 1.0"/>
```

(URDF)

render



LLM → 3D application via API

"Articulating desk lamp"

LLM

commands

```
import bpy
import random

def clear_scene():
    """Removes all objects from the
current scene to start fresh."""
    if bpy.context.object and
bpy.context.object.mode !=
'OBJECT':

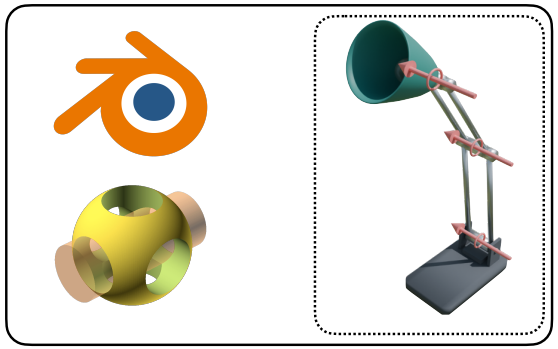
bpy.ops.object.mode_set(mode='O
BJECT')

    # Select and delete all mesh,
light, and camera objects

bpy.ops.object.select_all(action='DE
SELECT')
```

API

3D application

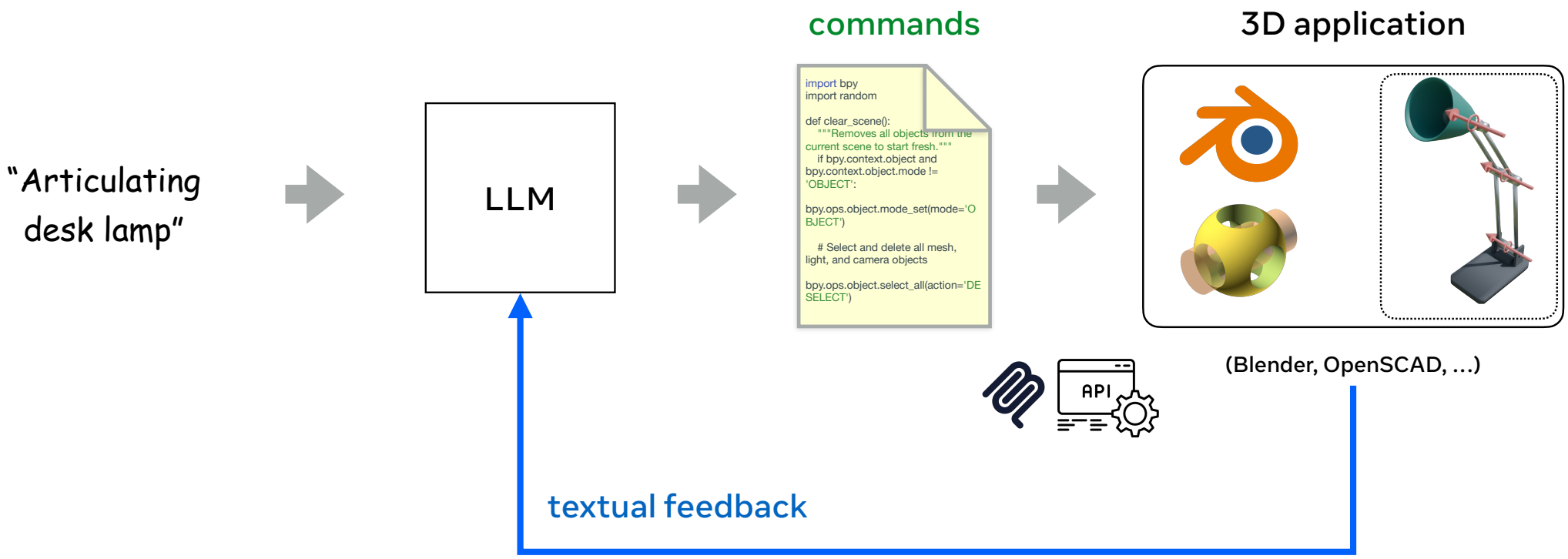


(Blender, OpenSCAD, …)

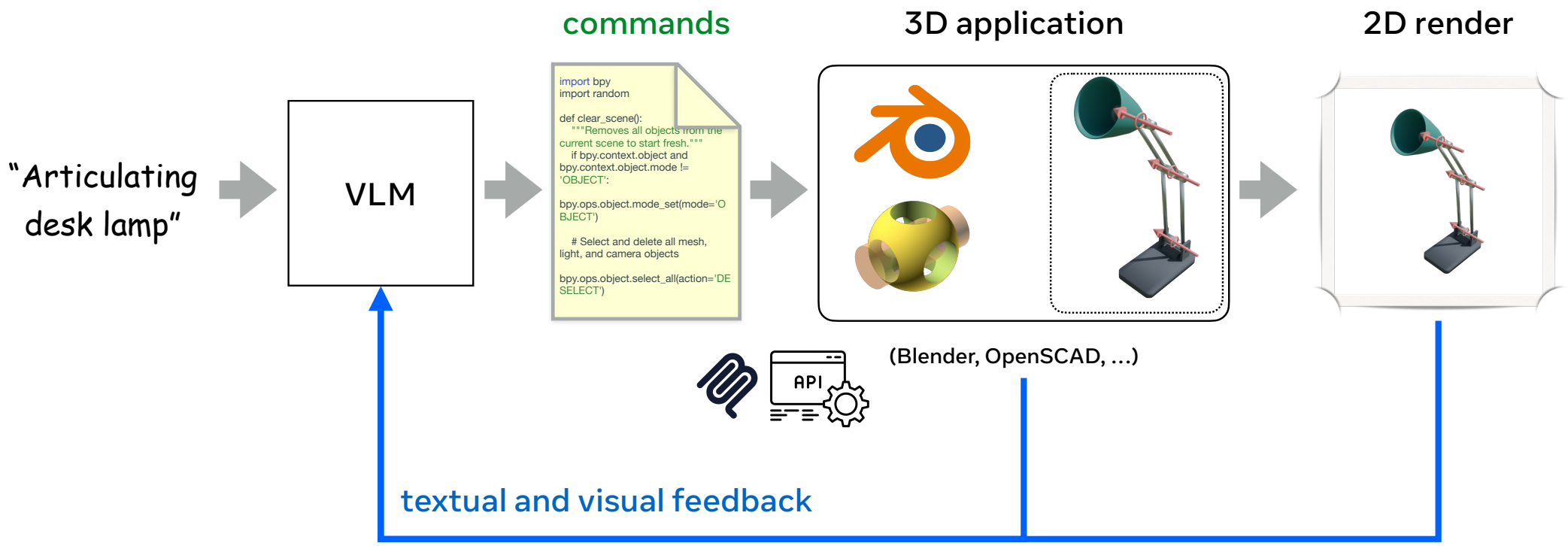
LLM

application

Feedback


"Articulating desk lamp"
LLM
commands
import bpy
import random
def clear_scene():
"""Removes all objects from the current scene to start fresh."""
if bpy.context.object and bpy.context.object.mode != 'OBJECT':
bpy.ops.object.mode_set(mode='OBJECT')
Select and delete all mesh, light, and camera objects
bpy.ops.object.select_all(action='DESELECT')
3D application
(Blender, OpenSCAD, ...)
API
textual feedback

Visual feedback



E.g., [**3D-Generalist.** Sun, et. al, 3DV, 2026] and [**VIGA.** Yin et al., arXiV, 2026].

Articraft

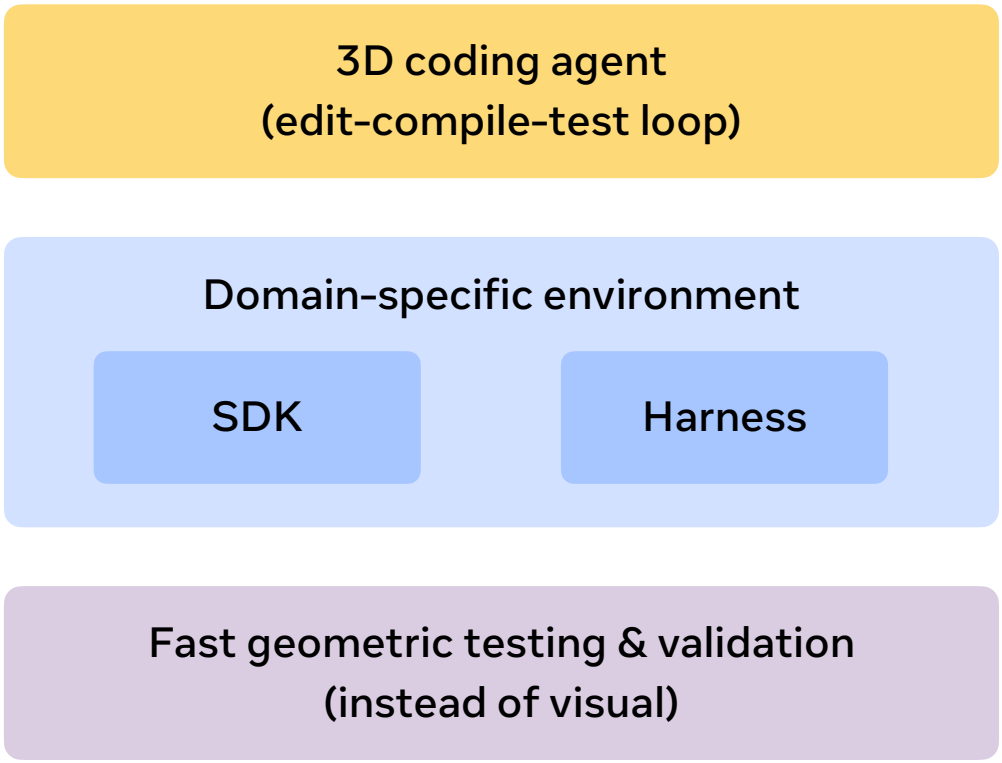
Goals

Construct **articulated 3D assets** from a single text prompt

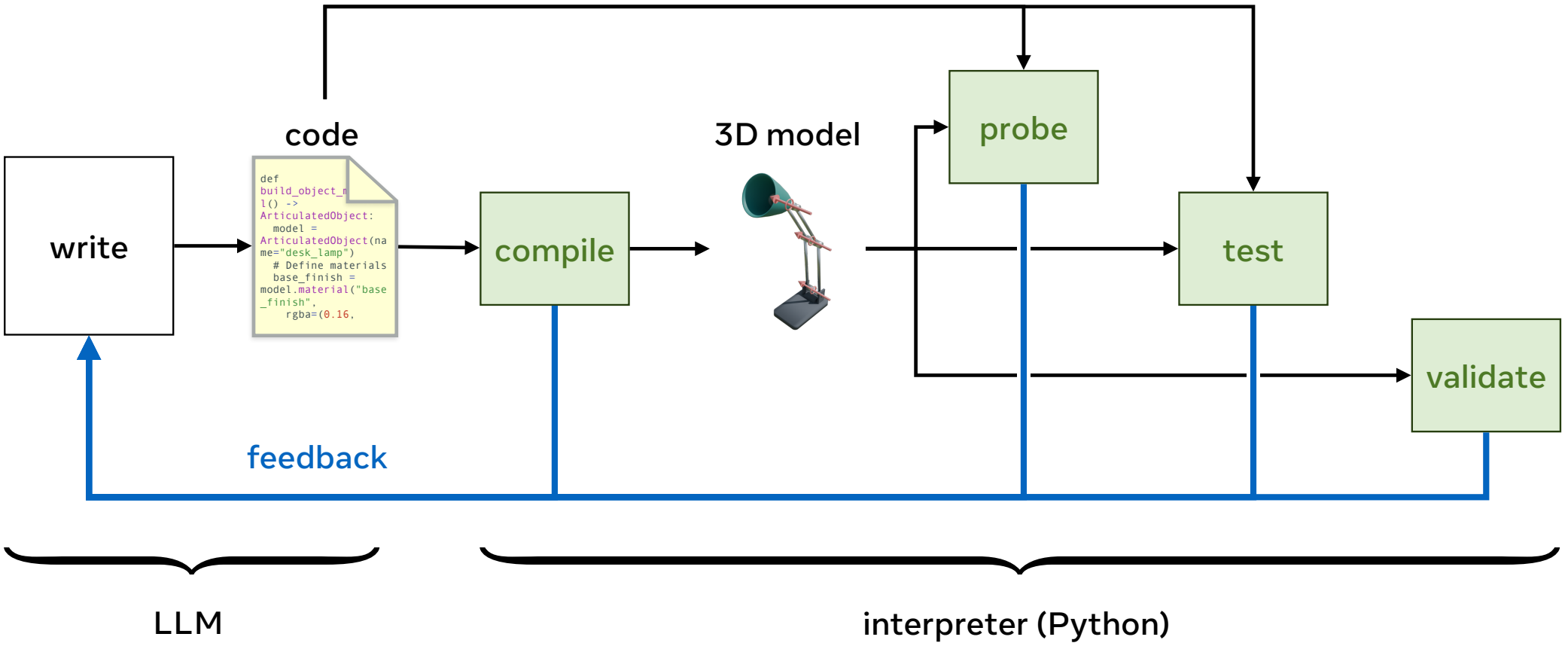
Use an off-the-shelf model (Codex, Claude, Gemini, …)

Fast and cheap (~$1 per asset)

How



3D coding agent



SDK

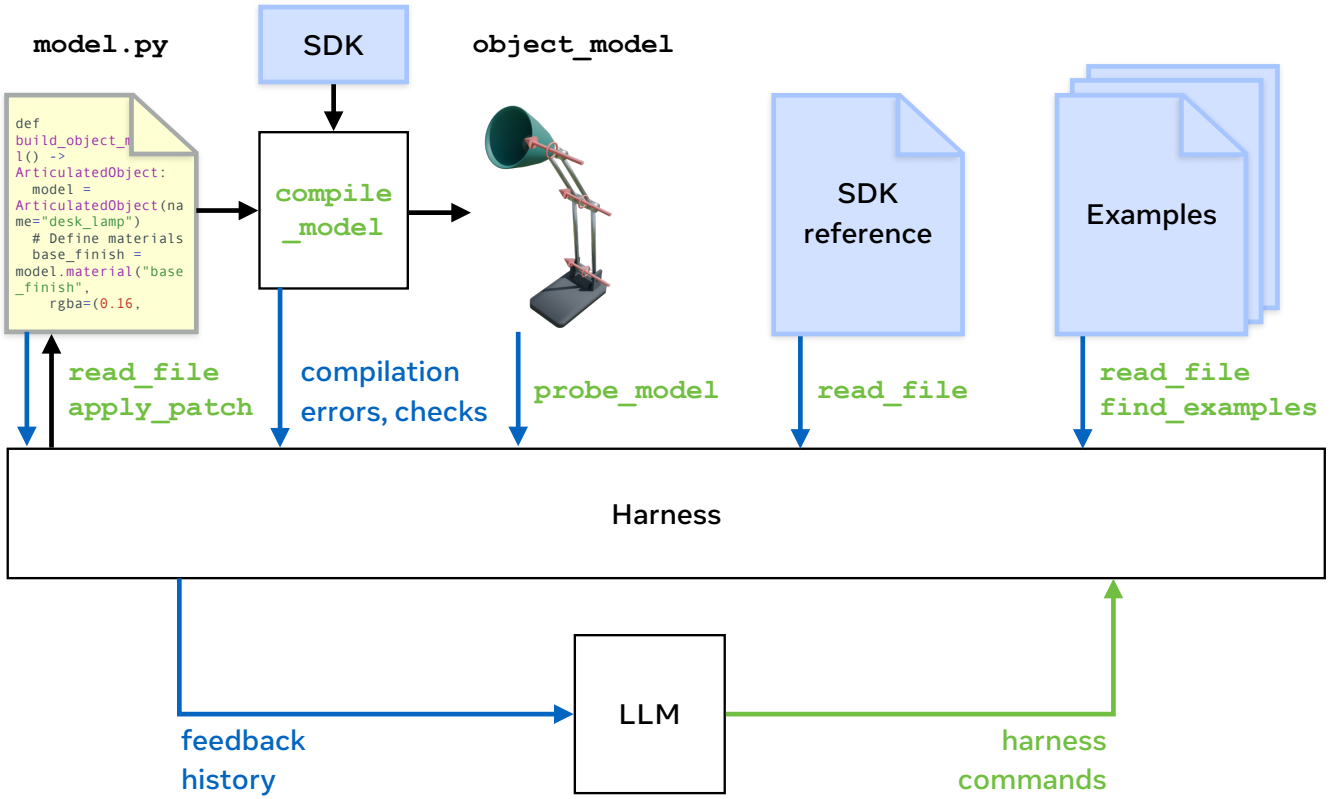
Construction

Low-to-high level primitives, operations:
`ArticulatedObject`, `Cylinder`, CadQuery, `PerforatedPanelGeometry`, ...

Testing

Express geometric tests, exceptions:
`expect_contact`, `expect_within`, ...

Harness



State

- `model.py`
- feedback

Read-only data

- SDK
- SDK reference
- SDK examples

Commands

- Read and write code
- Compile code
- Probe 3D model
- Read SDK reference
- Read SDK examples

Bicycle Crankset and Pedal Assembly

reasoning compressed into tool turns

STEP 01 / 27

PROMPT

Asset Brief

BRIEF

Workspace Documentation (read-only) The virtual workspace exposes `model.py` as the editable artifact script and `docs/` as

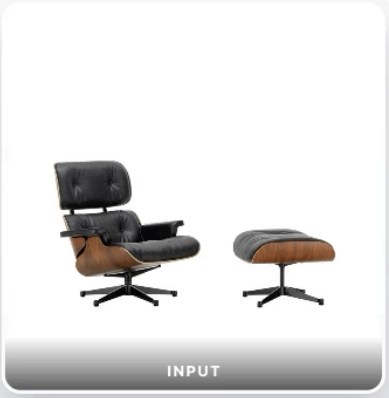
Workspace Documentation (read-only) The virtual workspace exposes `model.py` as the editable artifact script and `docs/` as read-only SDK guidance. `docs/sdk/references/quickstart.md` is the preloaded SDK

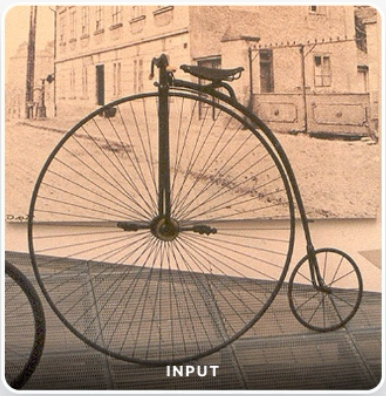
AWAITING RENDER

NO COMPILED ASSET YET

agent is reading context and preparing the first model revision


INPUT


INPUT

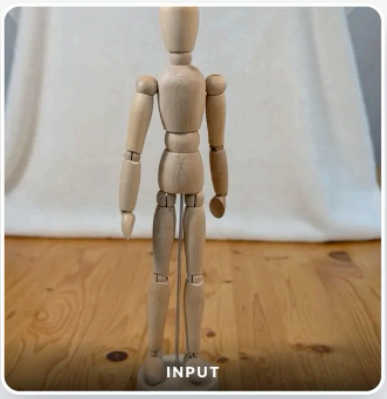

INPUT

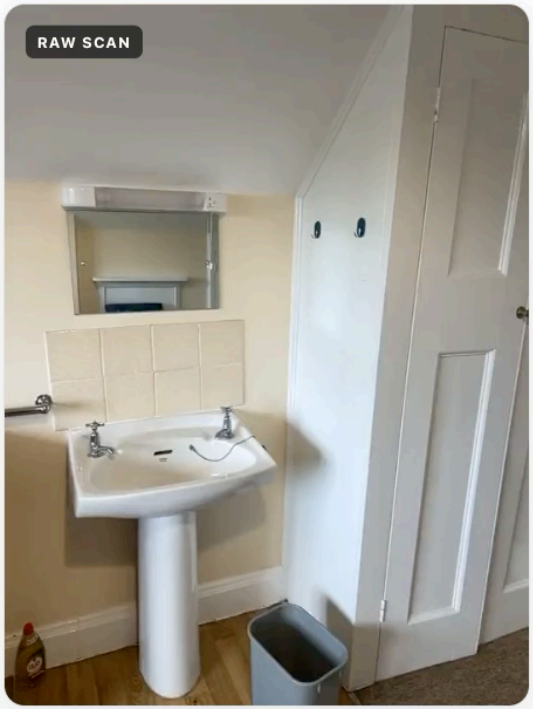

Nintendo GAME BOY
INPUT


INPUT

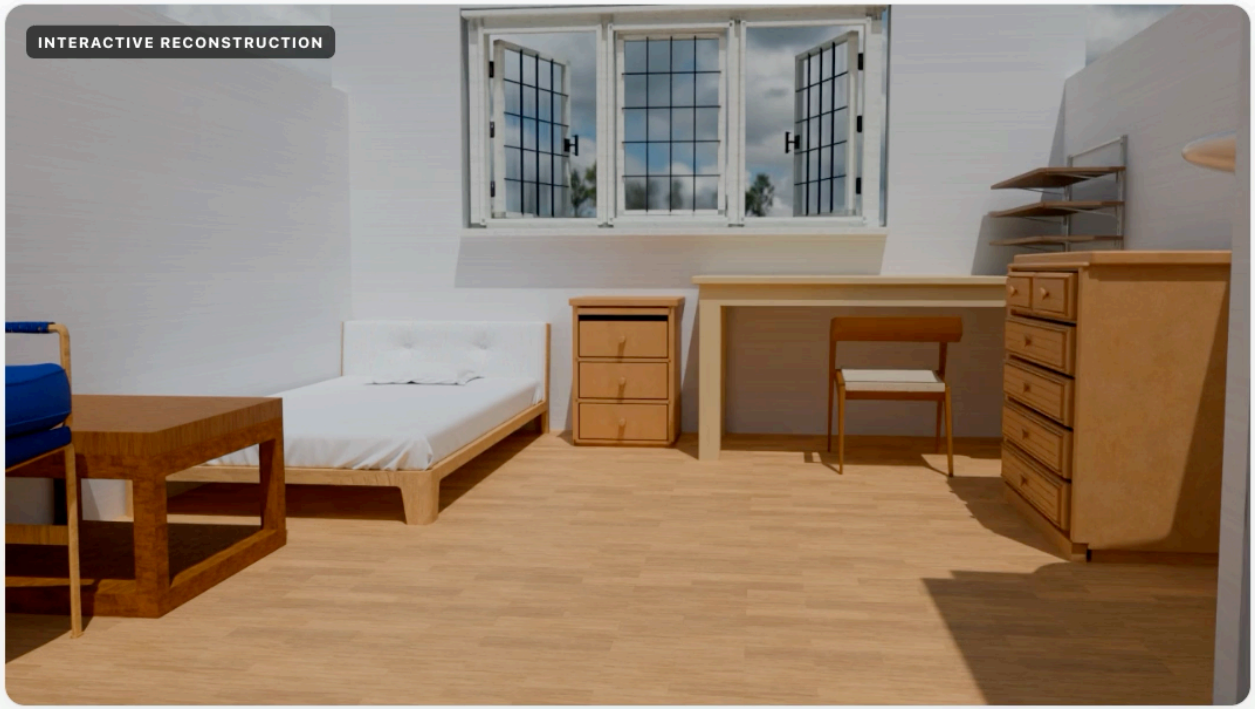

INPUT


INPUT


INPUT



LiteReality
+ Articraft









Articraft-10k

Articraft-10K: 10,000+ Articulated 3D Assets

SUPERCATEGORIES

10,764 mapped of 10,776 records

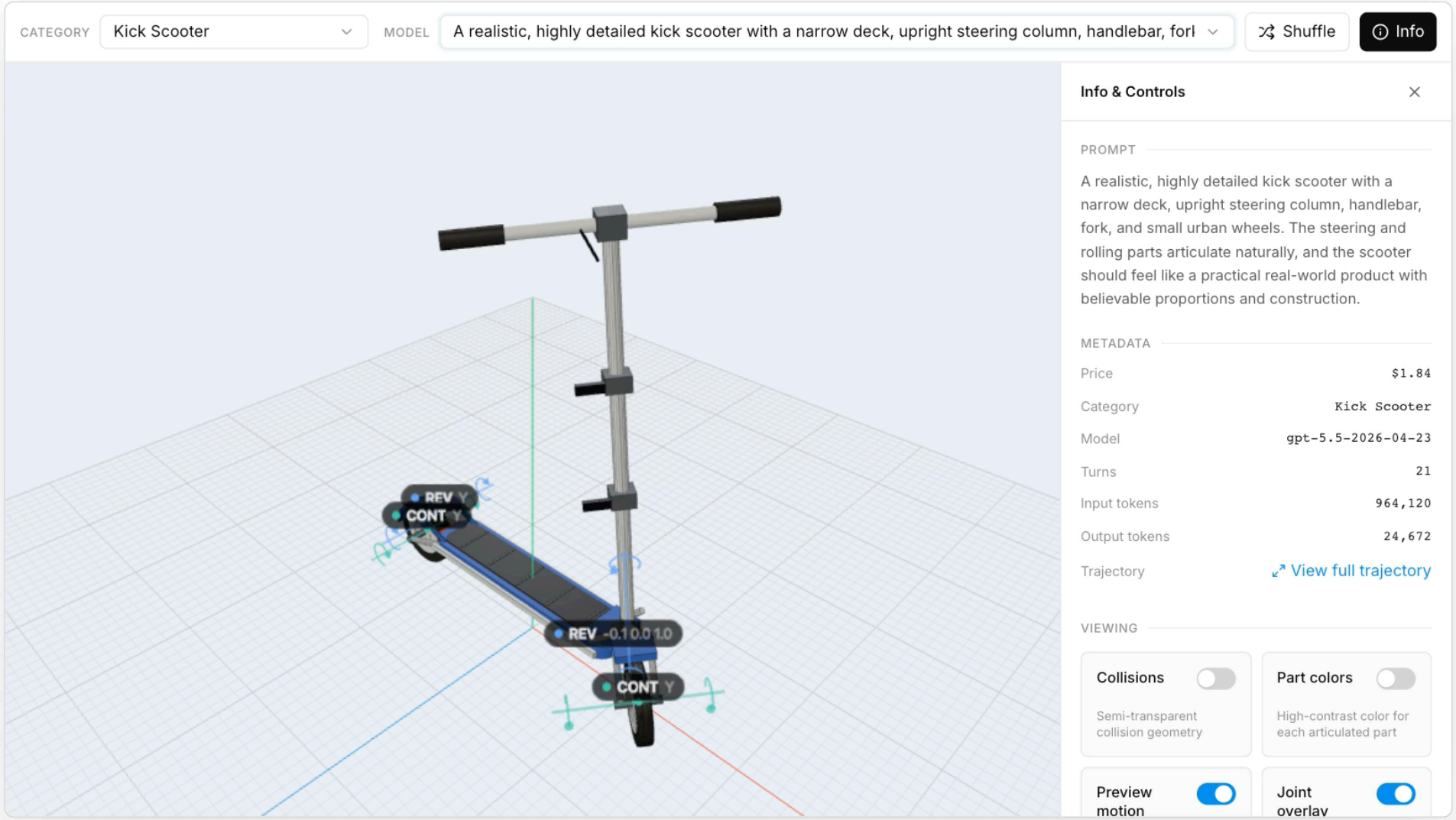
Showing 10,764 mapped records across 238 categories. 12 records in 5 categories are not assigned to a supercategory yet.



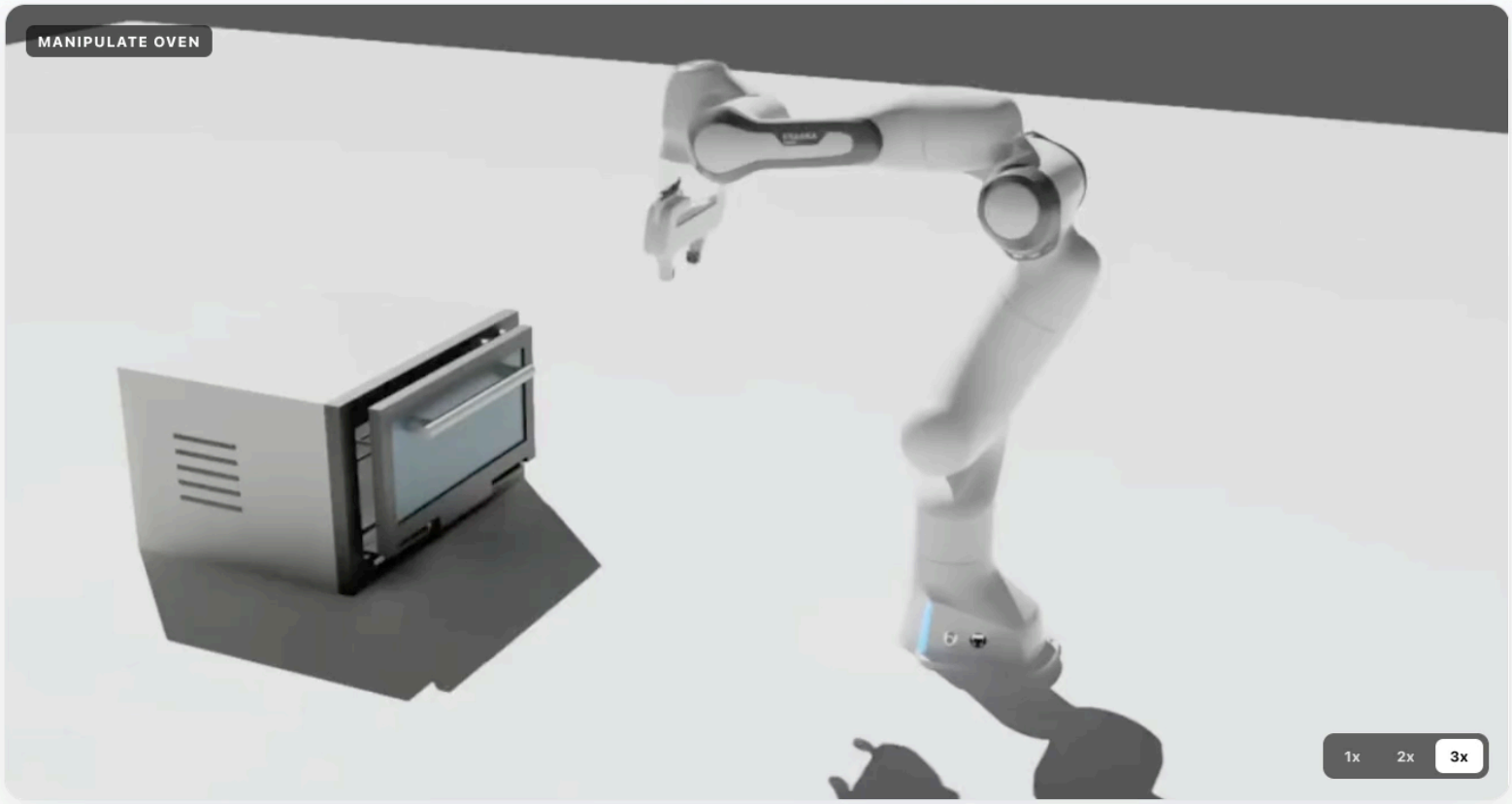
Distribution across mapped supercategories

GROUP		CAT	REC	STARS	COST	INPUT	OUTPUT
Primitive Joints & Modules	5.5%	10	591	4.7	$1.05	853k	27k
Kinematic Chains	9.0%	22	969	4.8	$1.11	754k	28k
Linear Stages & Gantries	4.8%	11	520	4.9	$0.896	577k	22k
Robotic Arms & Manipulators	5.8%	14	626	4.8	$1.02	713k	26k
Household Appliances	15%	36	1646	4.4	$1.24	957k	29k
Furniture & Storage	10%	22	1130	4.7	$1.06	729k	26k
Electronics & Optics	11%	27	1147	4.5	$1.10	755k	25k
Vehicles & Mobility	9.4%	24	1007	4.5	$1.36	978k	30k
Mounts & Positioning	4.7%	10	507	4.7	$1.16	824k	28k
Tools & Workshop	6.8%	19	737	4.3	$1.21	965k	28k
Doors, Gates & Access	5.8%	13	624	4.8	$0.979	588k	23k
Buildings & Infrastructure	5.2%	12	565	4.7	$1.11	725k	27k
Outdoor & Recreation	4.0%	11	433	4.6	$1.32	908k	28k
Military & Defense	1.2%	3	133	4.7	$1.27	857k	28k
Commercial Coin-Op & Vending	1.2%	4	129	4.2	$1.23	966k	28k

Interactive Viewer






MANIPULATE OVEN
1x
2x
3x

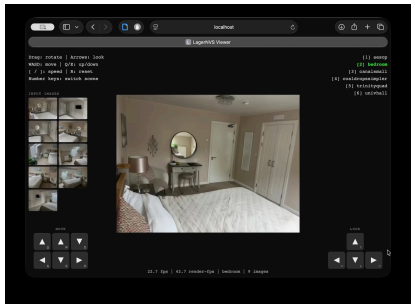

Manipulate Oven


Manipulate Drawer

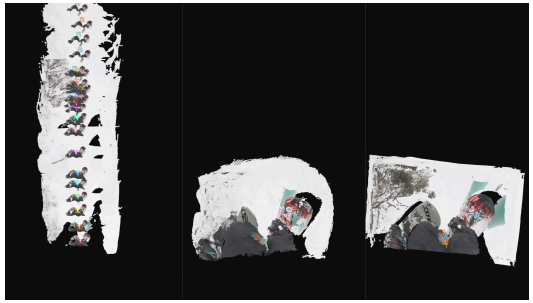

VR DEMO

Building a 3D foundation

3D transformers: power



3D transformers: scale



VGGT-Ω

3D construction



Articraft