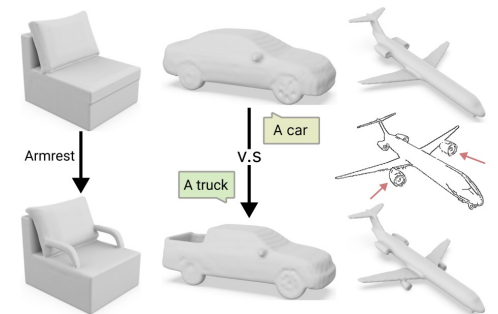


# Why is 3D Generation Hard?

And what can we do about it?



**CMPT464/764: Geometric Modeling in Computer Graphics**

*Lecture 13*

# Amazing image generation via LFM



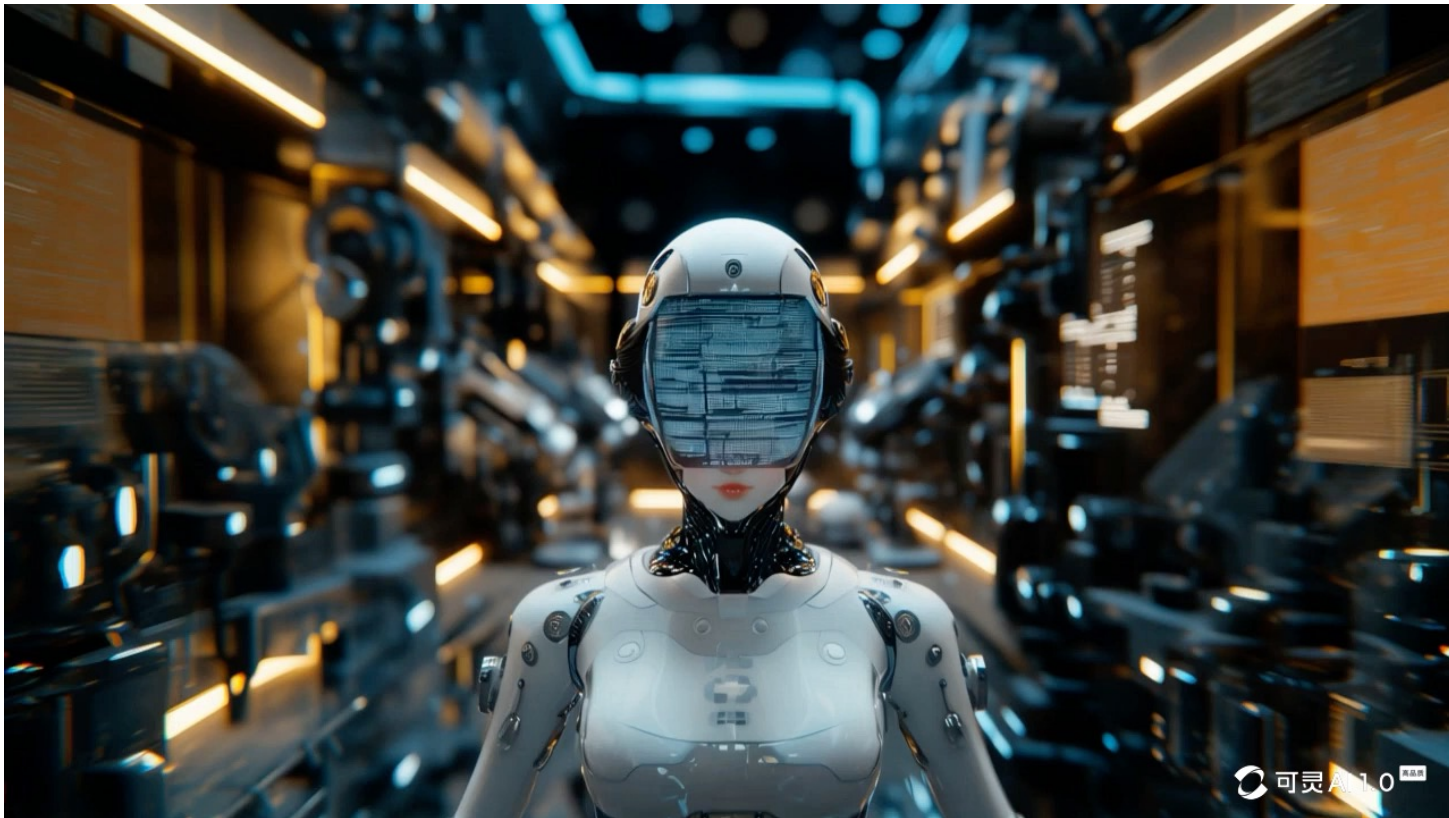
DALL-E 3

An illustration of an avocado sitting in a therapist's chair, saying 'I just feel so empty inside' with a pit-sized hole in its center. The therapist, a spoon, scribbles notes.

Amazon Titan Image Generator  
Change



# Video generation too



# State-of-the-art 3D generation

**RealmDreamer: Text-Driven 3D Scene Generation with Inpainting and Depth Diffusion**

Jaidev Shriram, Alex Trevithick, Lingjie Liu, Ravi Ramamoorthi

April 10, 2024

*"A bear sitting in a classroom with a hat on, realistic, 4k image, high detail"*





# Compare 2D ...



*“A bunny mesh”*

Image generated by DALL·E 3

# Compare 2D to 3D = geometry + texture



*"A bunny mesh"*

Image generated by DALL·E 3



*"A baby bunny sitting on top of a stack of pancakes"*

3D model generated from text [Zhu et al. 2023]

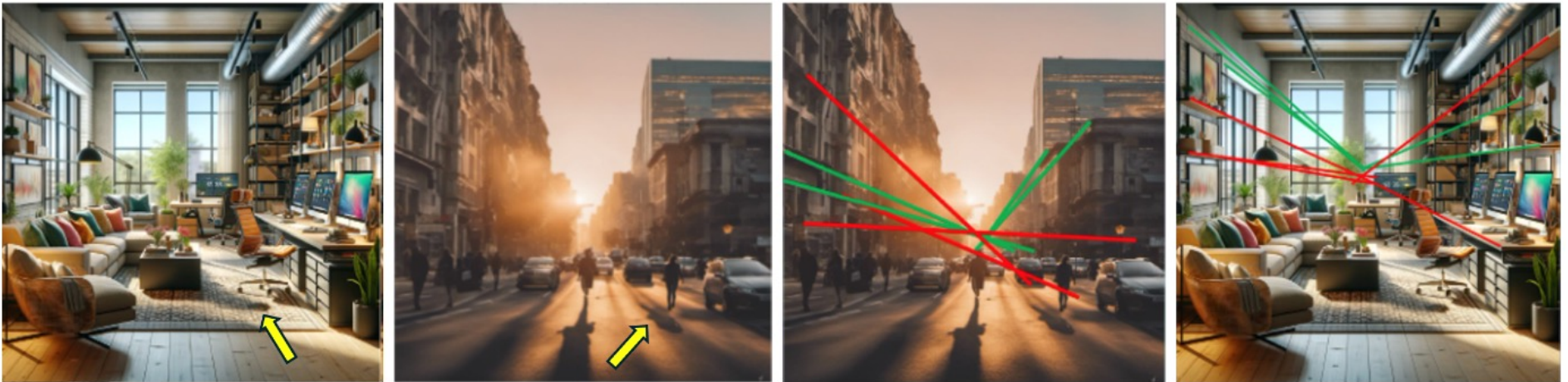


# What is geometry?

*Geometry is a branch of mathematics concerned with the properties of space such as the [distance](#), [shape](#), [size](#), and [relative position](#) of figures.*

Wikipedia

# Do LFM “understand” (projective) geometry?



Images generated by [Stable Diffusion](#)

“Shadows do not lie and lines do not bend ...” [Sarkar et al. CVPR 2024]



# More obvious error: “AI hands”



"Human hands" generated by Stable Diffusion (left), DALL-E 2 (upper right), and Midjourney (lower right)

# How come?



"Human hands" generated by Stable Diffusion (left), DALL-E 2 (upper right), and Midjourney (lower right)

- Hands in an image are
- ❖ usually small
  - ❖ often obscured/occluded
  - ❖ inconsistent in appearance
  - ❖ Human hand having five fingers is an **obvious fact**: likely not in the captions even if the images do involve human hands

# LFMs are good at learning ...

😊 consistent and prominent features, e.g., dogs, but not dog ears

# LFMs are good at learning ...

- 😊 consistent and prominent features, e.g., dogs, but not dog ears
- 😄 in a “zero-shot” way to generalize w/o laborious labeling



# LFMs are *not* so good at ...

- 😊 learning consistent and prominent features: dogs vs. dog ears
- 😊 in a “zero-shot” way to generalize w/o laborious labeling
- 😞 discovering **obvious, but “hidden”, rules/laws/concepts**: e.g., not in training data (captions), highly abstract, hard to extract ...

$$\vec{a} = \frac{\Sigma \vec{F}}{m} = \frac{\vec{F}_{net}}{m}$$
$$\frac{m}{s^2} = \frac{N}{kg}$$

Newton's Laws



Shadow formation



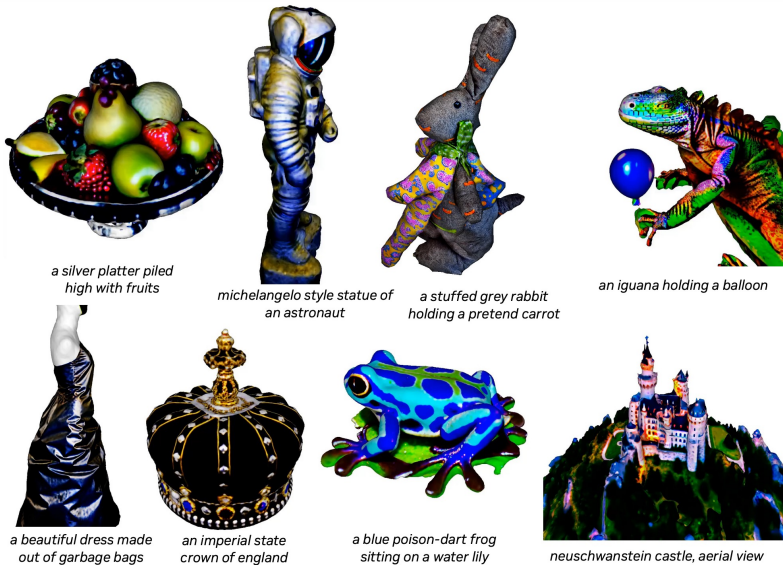
Vanishing points



How chairs stand

# Correct geometry hard for 3D GenAI

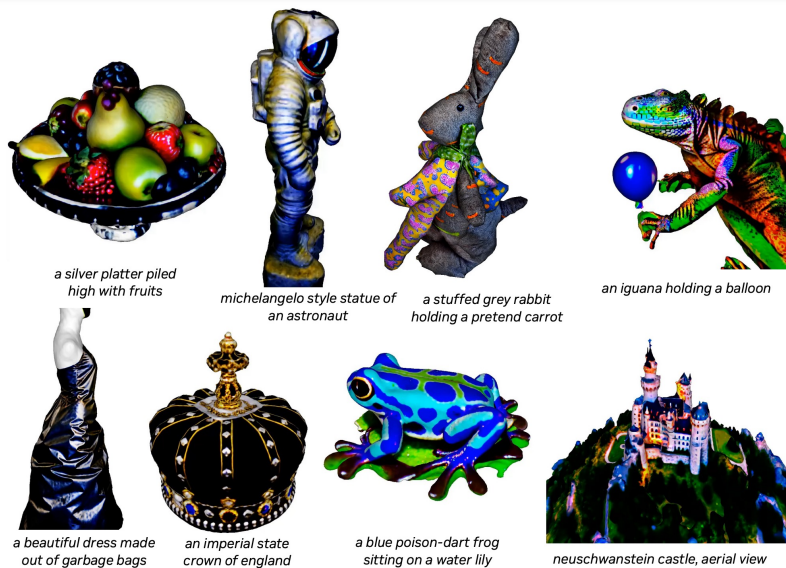
❖ Incidence, size, coplanarity, symmetry, relative position, etc.



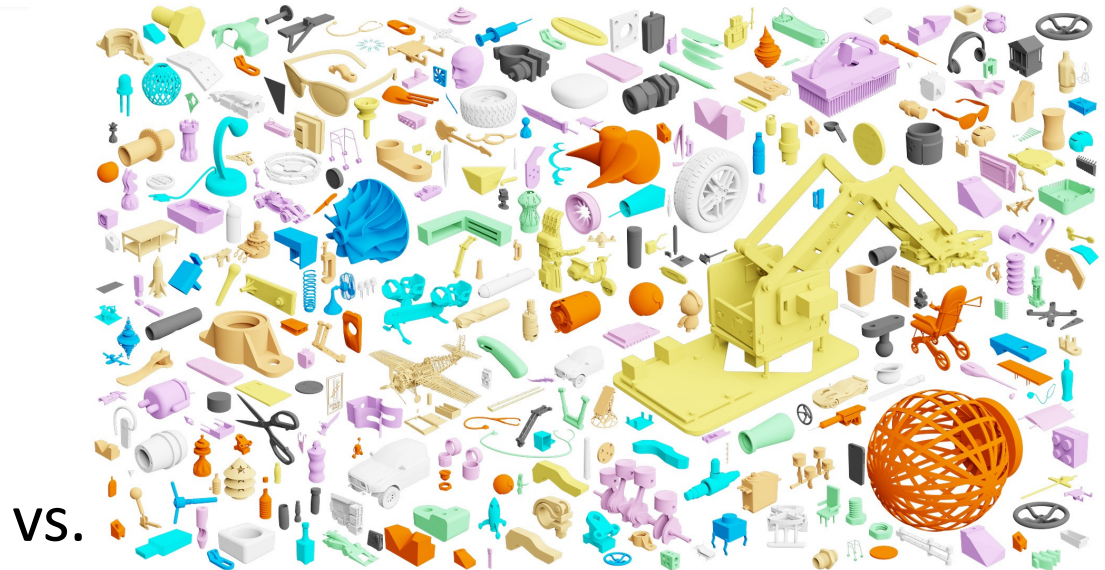
Contemporary 3D GenAI results (e.g., from Magic3D) are for “viewing pleasure”?

# Correct geometry hard for 3D GenAI

❖ Incidence, size, coplanarity, symmetry, relative position, etc.



Contemporary 3D GenAI results (e.g., from Magic3D) are for “viewing pleasure”?



3D objects and designs meant to be used:  
from Autodesk Fusion 360 dataset

# LLMs “know” geometry ... in texts

ChatGPT 4



How big is a computer mouse?



The size of a computer mouse can vary significantly based on the type and design. Generally, most standard computer mice are about 4 to 5 inches (10 to 12.5 cm) in length and 2 to 3 inches (5 to 7.5 cm) in width. Some ergonomic mice may be larger or shaped differently to fit the hand better, while compact mice designed for portability can be smaller, around 3 to 4 inches (7.5 to 10 cm) in length. If you're looking for a specific model, it's best to check the manufacturer's specifications.



# LLMs “know” geometry ... in texts

Claude 3 Sonnet



You

What is a vanishing point?



**AI** Claude



A vanishing point is a concept used in linear perspective drawing and painting. It refers to the point on the horizon line where parallel lines appear to converge or "vanish" in the distance.

Some key points about vanishing points:

- They are used to create the illusion of depth and three-dimensionality on a two-dimensional surface like a canvas or paper.
- Parallel lines that are perpendicular to the viewing plane (like railroad tracks) will have a single vanishing point on the horizon line that they appear to converge towards.
- Parallel lines that are angled relative to the viewing plane (like the corners of buildings) will have two vanishing points on the horizon line.
- The horizon line represents the eye level of the viewer. Objects above it appear higher, below it appear lower.
- The closer an object is to the vanishing point(s), the smaller it appears due to the illusion of distance.

Accurately plotting vanishing points allows artists to construct proper perspective and create realistic spatial depth in their drawings and paintings. Mastering vanishing points is a key skill in linear perspective techniques.

# Visual manifestation a different story

3D generations by Tripo3D



*"A real mouse"*

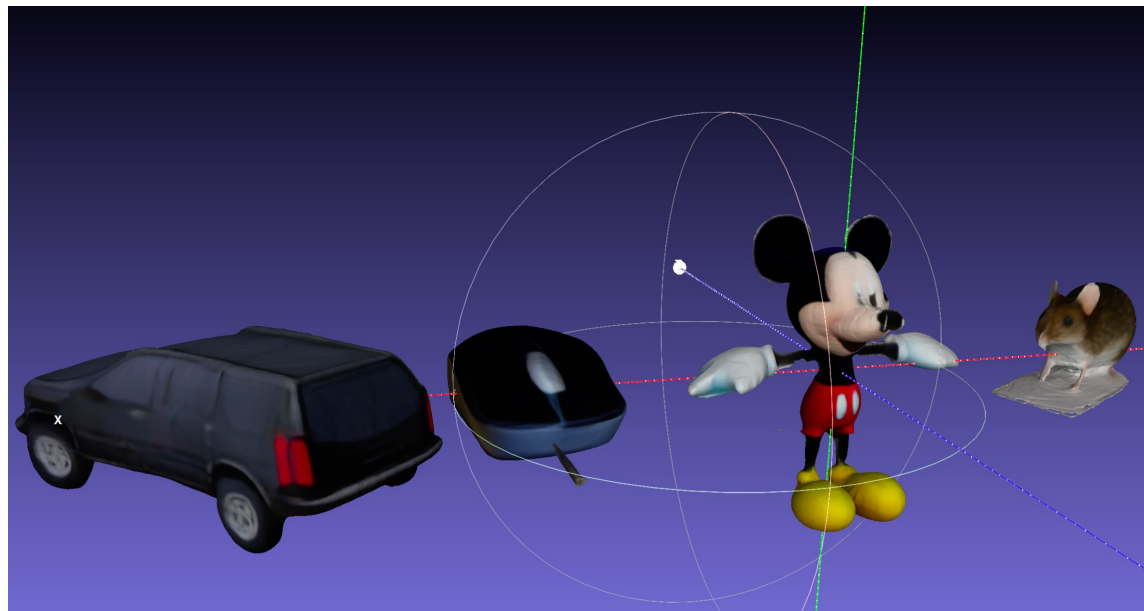


*"A computer mouse"*



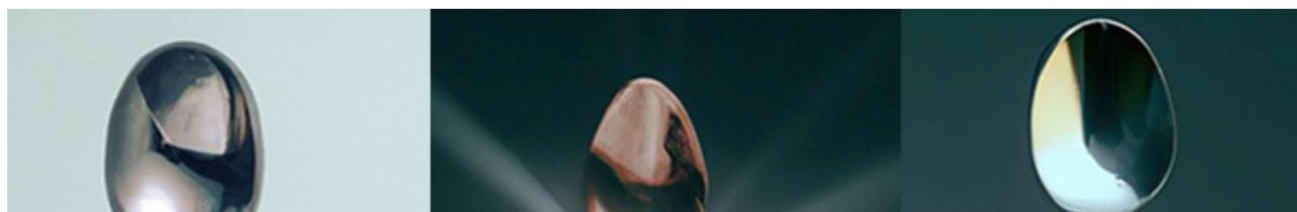
*"Mickey mouse  
in Magic Kingdom"*

# 3D generation by One-2-3-45++



- ❖ All the generated 3D models are simply **normalized**
- ❖ Contemporary 3D GenAI models do not have a proper sense of size

## Another example: spatial relations



*“DALL-E 2’s difficulty with even basic **spatial relations** (such as in, on, under) suggests that whatever it has learned, it has not yet learned the kinds of representations that allow humans to so flexibly and robustly **structure** the world.”*

*“A direct interpretation of this difficulty is that systems like DALL-E 2 do not yet have **relational compositionality**.”*

*From “Testing Relational Understanding in Text-Guided Image Generation”, by Colin Conwell and Tomer D. Ullman, 2022*

DALL-E 2 results on “A Cup on a Spoon”



# For 3D scene generation

## Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes

Dana Cohen-Bar

Elad Richardson

Gal Metzer

Raja Giryes

Daniel Cohen-Or

- ❖ Manually create 3D scene proxy with size and spatial structures



# For 3D scene generation

## Set-the-Scene: Global-Local Training for Generating Controllable NeRF Scenes

Dana Cohen-Bar

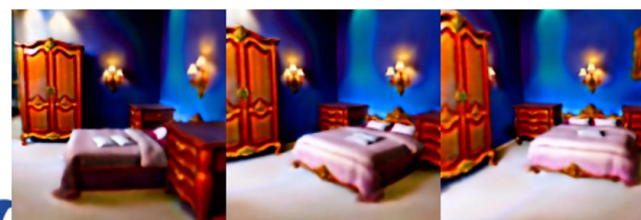
Elad Richardson

Gal Metzer

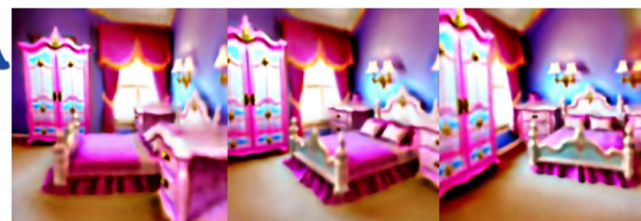
Raja Giryes

Daniel Cohen-Or

- ❖ Manually create 3D scene proxy with size and spatial structures
- ❖ Then generate **per-object NeRFs**



"A Baroque Bedroom"



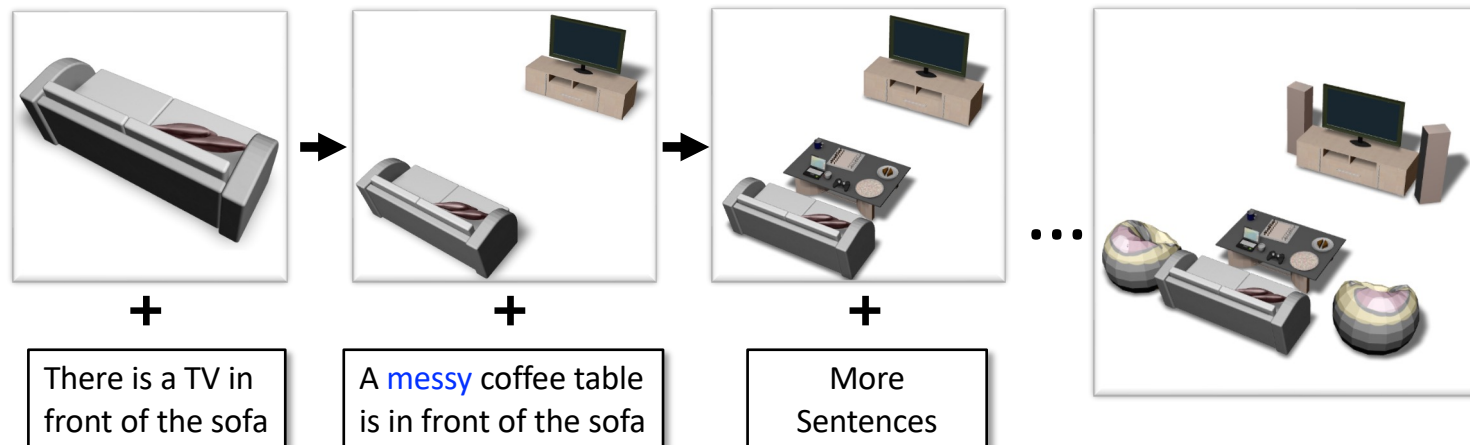
"A Princess Bedroom"

# Old, pre-DL work (2018)

## Language-Driven Synthesis of 3D Scenes from Scene Databases

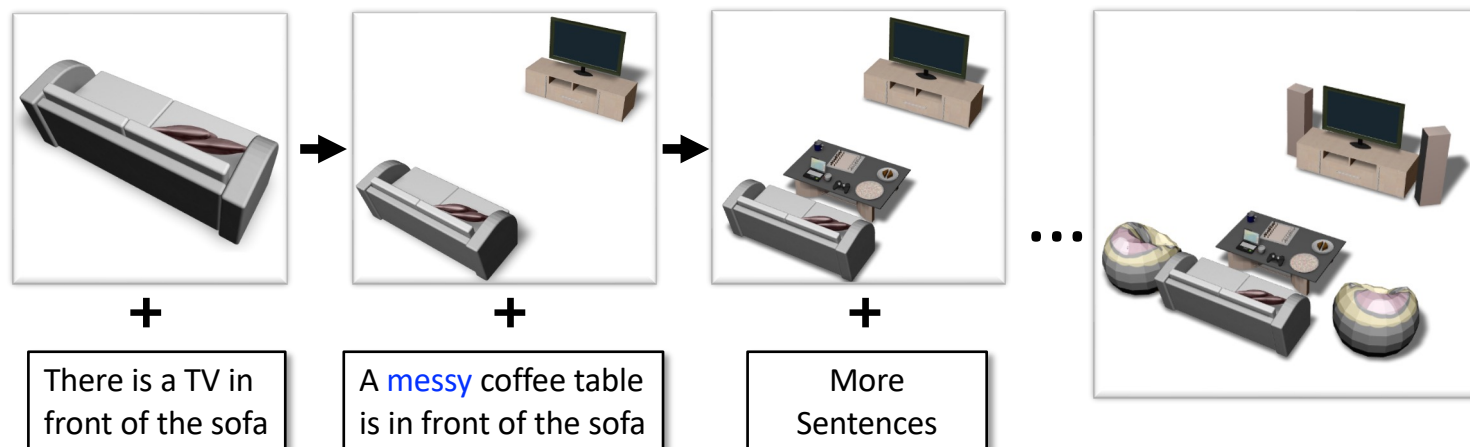
Rui Ma\* <sup>1,2</sup> Akshay Gadi Patil\* <sup>1</sup> Matthew Fisher <sup>3</sup> Manyi Li <sup>4,1</sup> Sören Pirk <sup>5</sup>  
Binh-Son Hua <sup>6</sup> Sai-Kit Yeung <sup>7</sup> Xin Tong <sup>8</sup> Leonidas Guibas <sup>5</sup> Hao Zhang <sup>1</sup>

[Ma et al. SIGGRAPH Asia 2018]



# Automate layout generation

- ❖ Supervised learning: spatially and semantically aware annotations
- ❖ Automatic scene layout generation from natural language commands
- ❖ Subscene/object retrieval and accommodation, e.g., TV  $\rightarrow$  TV + TV stand

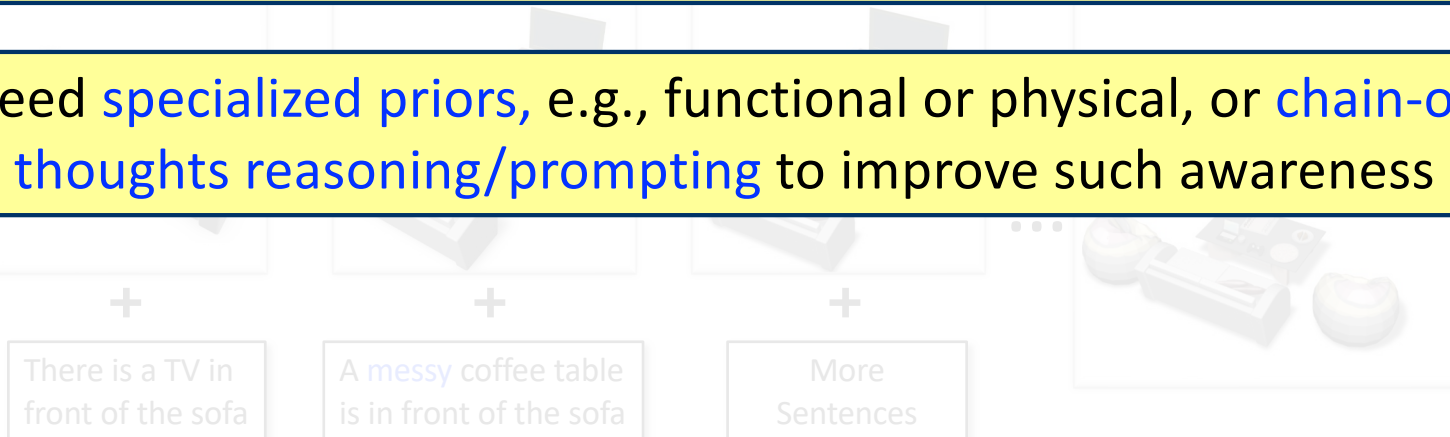


# Automate layout generation

- ❖ Supervised learning: spatially and semantically aware annotations
- ❖ Automatic scene layout generation from natural language commands

Most zero-shot Text-to-X models **do not have clear spatial or geometry awareness**, e.g., region localization, object orientation, relations, etc.

Need **specialized priors**, e.g., functional or physical, or **chain-of-thoughts reasoning/prompting** to improve such awareness

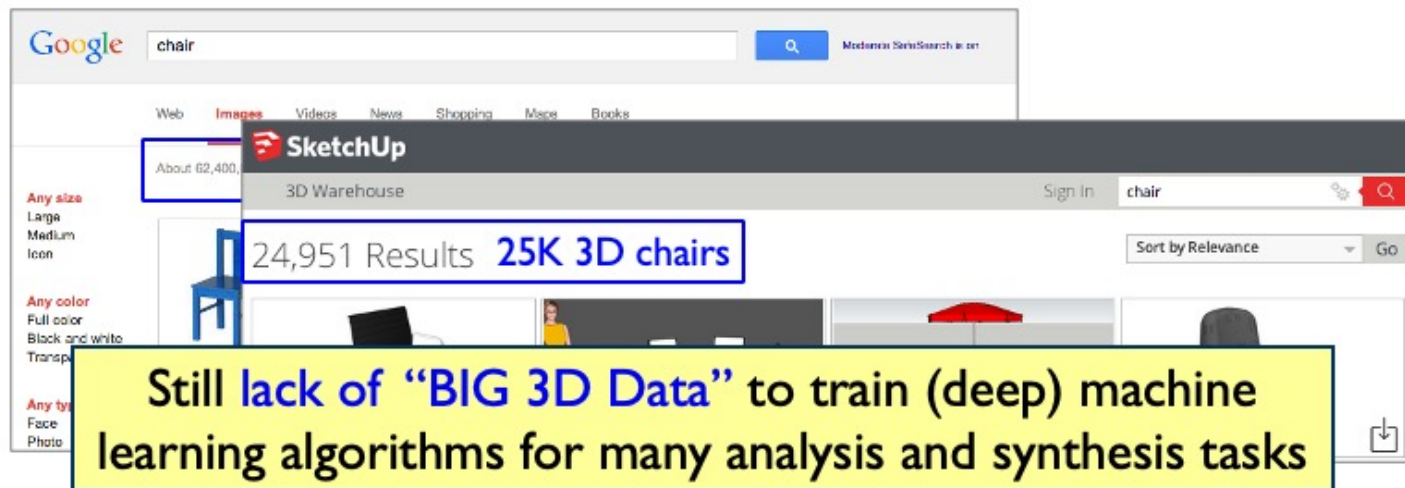




# Next challenges: 3D data challenge

## Last lecture

- Acquisition of and interaction with 3D contents are hard



## #2: 3D data challenge in 2024

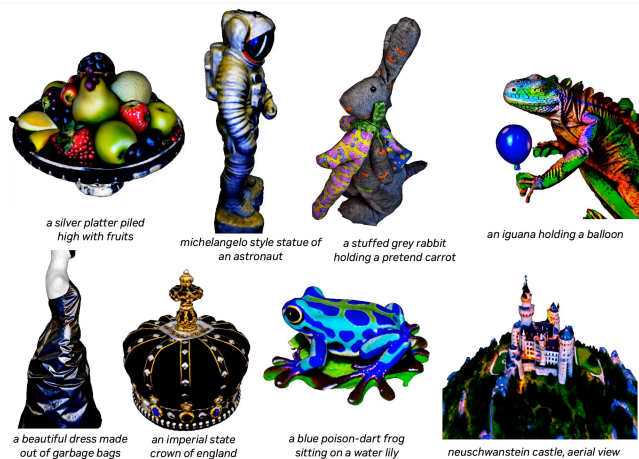
- ❖ Gap between 3D models and images/texts has **only widened!**

# #3: 3D output challenge

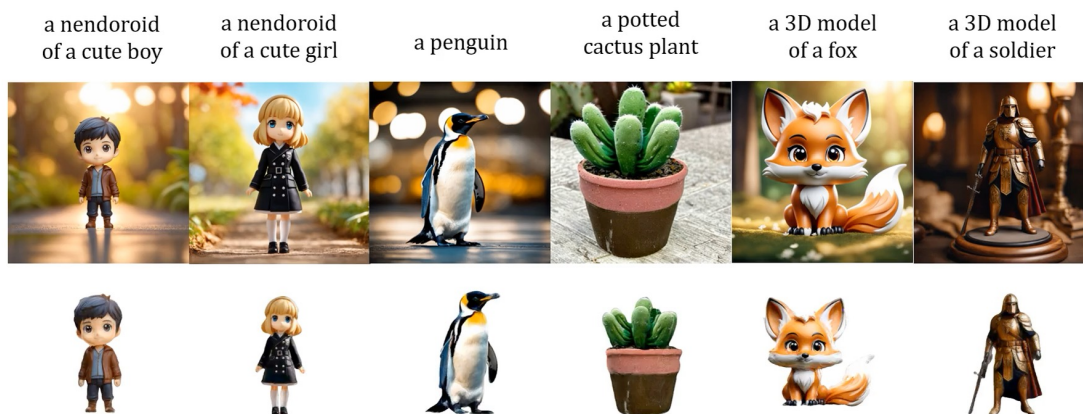
Must go beyond appearance ( $\Leftarrow$  differentiable rendering).

Think about **functionality!**

Ultimately, generated 3D shapes should **serve intended functions!**



Magic3D

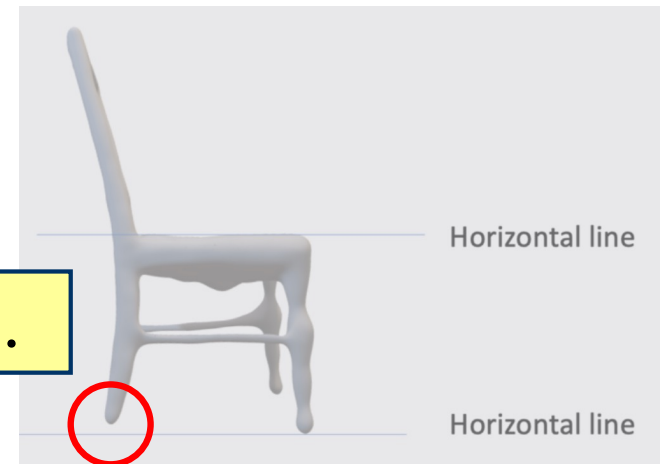
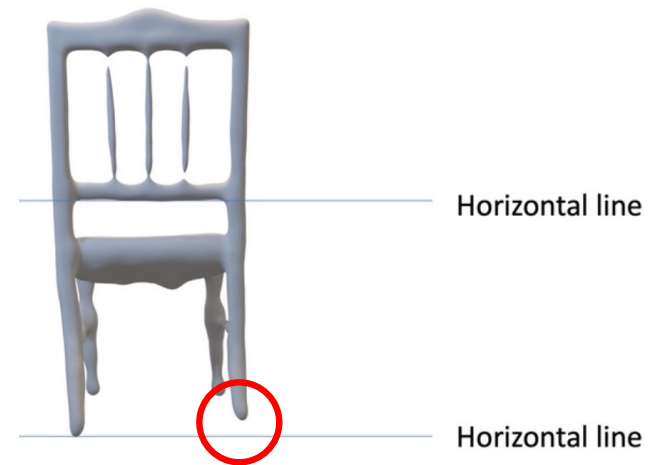


DreamGaussian

# SOTA 3D generation [2024]



Multi-view  
consistent synthesis  
from single image



Consistency does not imply correctness ...

## #4: Controllability challenge

- ❖ Amount of 3D models vs. images/texts – gap only widened!
- ❖ 3D generations should not just be looked at, e.g., for NVS; they ought to be used and interacted with, as in the real world
- ❖ Text/image conditions offer **no fine-grained control for editing**



## #4: Controllability challenge

- ❖ Amount of 3D models vs. images/texts – gap only widened!
- ❖ 3D generations should not just be looked at, e.g., for NVS; they ought to be used and interacted with, as in the real world
- ❖ Text/image conditions offer no fine-grained control for editing



Text2Scene is **too coarse** — manually create 3D scene proxy for generation and editing

[Cohen-Bar et al. ICCV 2023]

# Most fundamental: representation challenge

❖ Amount of 3D models vs. images/texts – gap only widened!

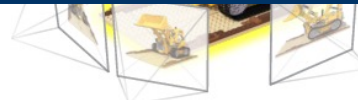
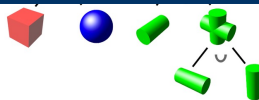
❖ 3D generations should not just be looked at as a success for NVC, they

“While NeRFs are often thought about as a major victory for AI-based techniques, ..., the **key insight** may actually simply be in the idea of a continuous volumetric **representation** (neural implicit **representation** from 2019) ... (NeRFs are) by no means the final word on the “best” 3D scene **representation** for **image-based rendering**.

Ravi Ramamoorthi

“NeRFs: The Search for the **Best 3D Representation**”, arXiv, August 2023

Representation is (almost) **everything!**

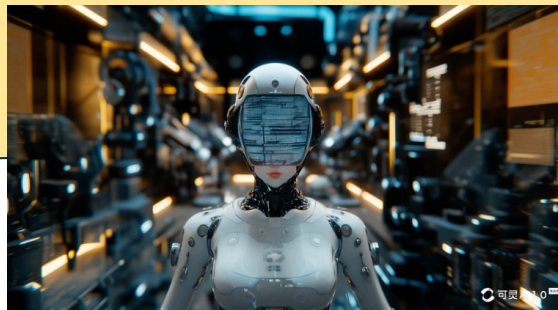
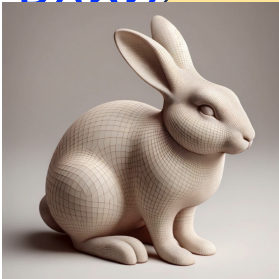


# Why is 3D generation hard?

- ❖ Data challenge: widening gap between 2D and images/texts
- ❖ Output of 3D models is often limited to static functions
- ❖ Control of editing
- ❖ Representation
- ❖ Series

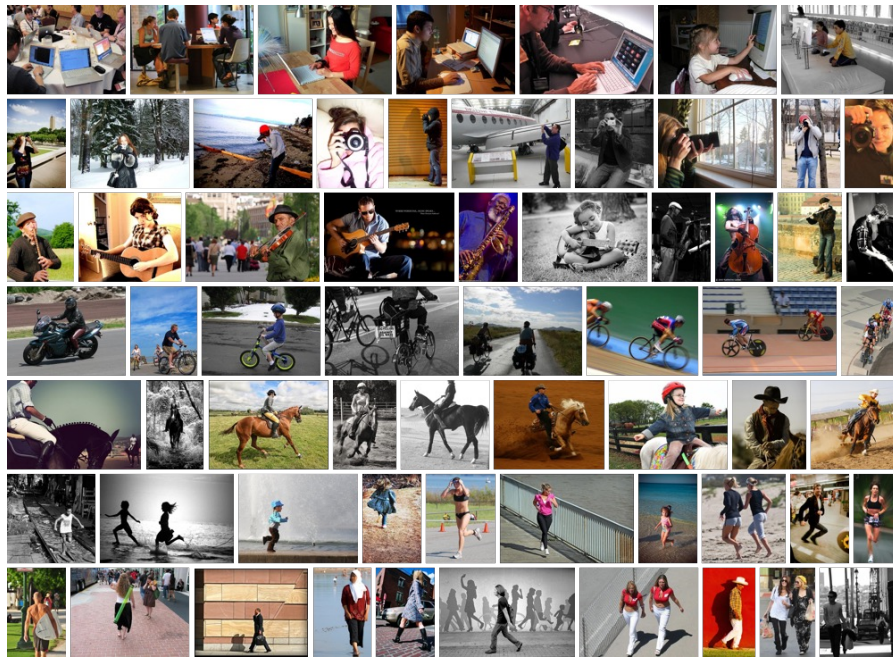
Do we need 3D at all?

(Just do video diffusion or flow matching)



# Observers vs. doers

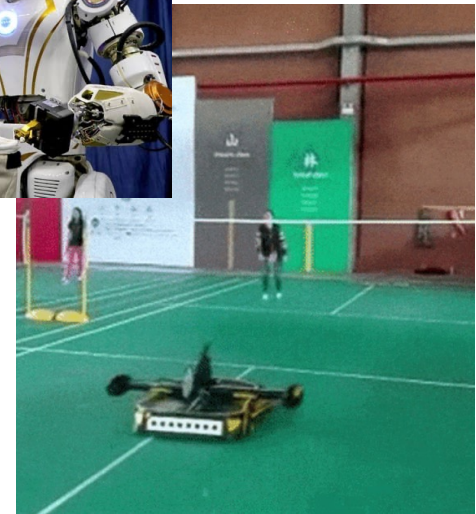
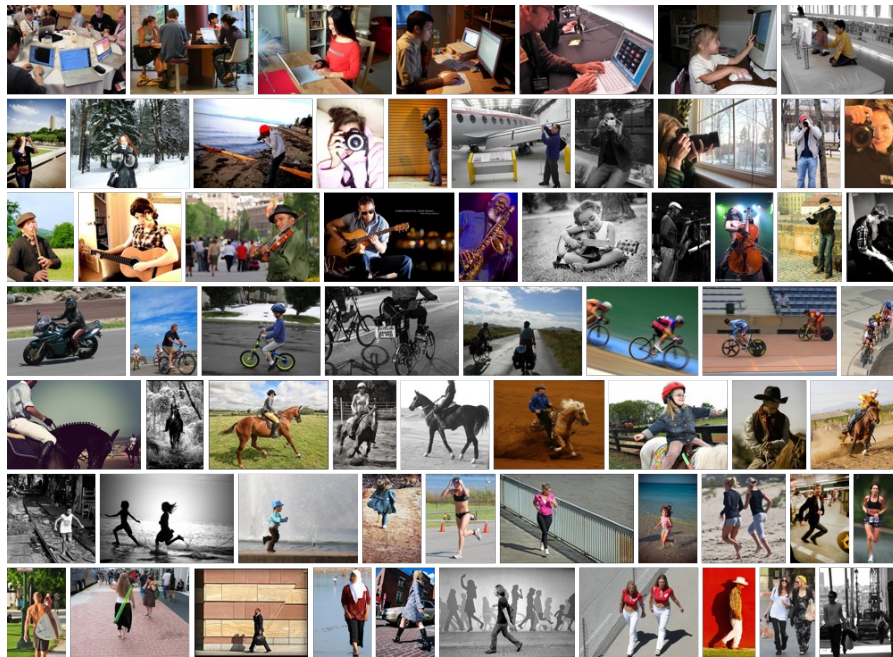
❖ We are not just passive **observers** — we are **doers**!





# Observers vs. doers

❖ We are not just passive **observers** — we are **doers**! **Robots** too!





# Observers vs. doers vs. creators

- ❖ We are not just passive observers — we are doers! Robots too!
- ❖ We are also 3D creators

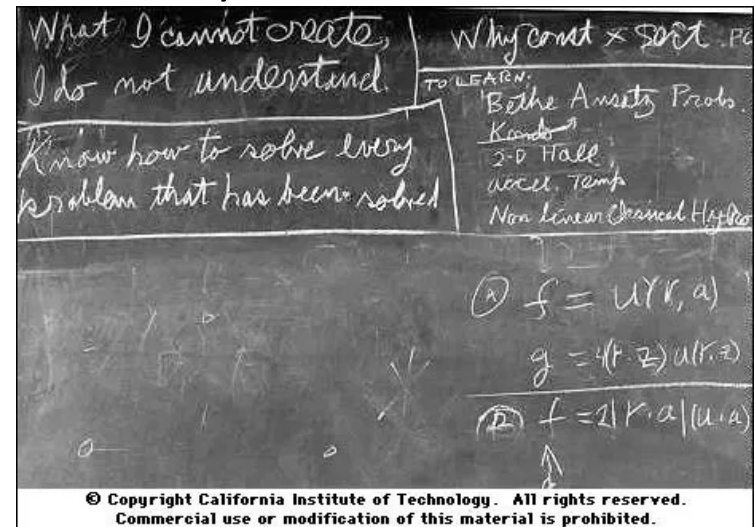


# Observers vs. doers vs. creators

- ❖ We are not just passive observers — we are doers! Robots too!
- ❖ We are also 3D creators — we create (generate) to understand!



Richard Feynman:



# To counter 3D data challenge

- ❖ 3D reconstruction and generation **without 3D supervision**





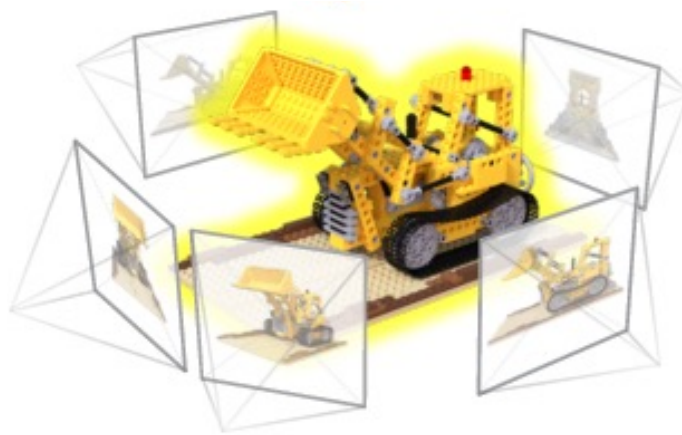
# Current trend

❖ Text-to-3D, image-to-3D, or text-to-image-to-3D

Text:  
*"A bunny on pancakes"*



Single-view image



NeRF [Mildenhall et al. ECCV 2020]  
via multi-view



Magic3D  
[Lin et al. CVPR 2023]

# Current trend

❖ Text-to-3D, image-to-3D, or text-to-image-to-3D

Text:  
*"A potted cactus plant"*



Single-view image



3D Gaussians via multi-view  
[Kerbl et al. SIGGRAPH 2023]



DreamGaussian  
[Tang et al. ICLR 2024]

# Critical issue #1

😞 NeRF and 3DGS produce **rendering, not modeling**, primitives



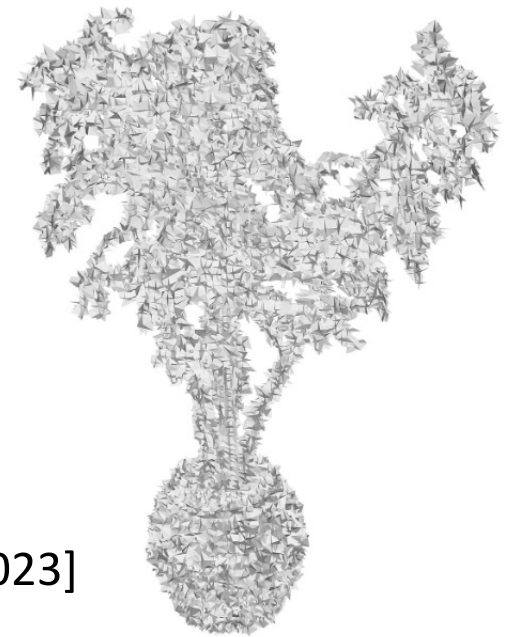
Images from MobileNeRF [Chen, Tagliasacchi et al. CVPR 2021]



# Critical issue #1

😞 NeRF and 3DGS produce **rendering, not modeling**, primitives

- **Unstructured**: not how human reasons about 3D
- **Not editable/reusable**: for further modeling
- **Not functional**: **true purpose** of 3D creations



Images from MobileNeRF [Chen, Tagliasacchi et al. CVPR 2023]

# Critical issue #2

☹️ NeRF and 3DGS produce rendering, not modeling, primitives

☹️ Avoiding 3D supervision is **unnatural**



vs.



Learning by **interacting in 3D**

# Critical issue #3

😞 NeRF/3DGS produces rendering, not modeling, primitives

😞 Avoiding 3D supervision is unnatural

😞 Avoiding 3D supervision leads to **VERY LARGE model**

- Training data: **billions of images** to train diffusion models
- Neural network size: **billions of network parameters**
- Training costs: **days/weeks** costing **millions of \$**, a privilege of large corporations — professors could only use their pre-trained models 😊
- Environmental impact: excessive use of **electricity and heat**

# Critical issue #3



Andrew Ng's data-centric AI: "smartizing" training data.  
"The key is good (quality) data, not big data."

Nothing wrong with using few 3D data, overall cost may be smaller!  
"One 3D model is worth a thousand images, maybe more!"  
"Supervised" is not always "inferior" to "unsupervised/self-supervised", what matters is the overall cost!

# An extreme example

- ❖ Train with a **single 3D model** while attaining **generalizability**

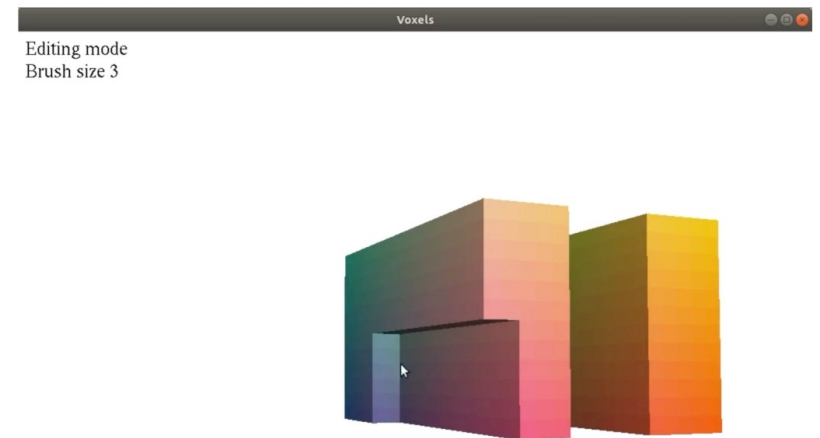
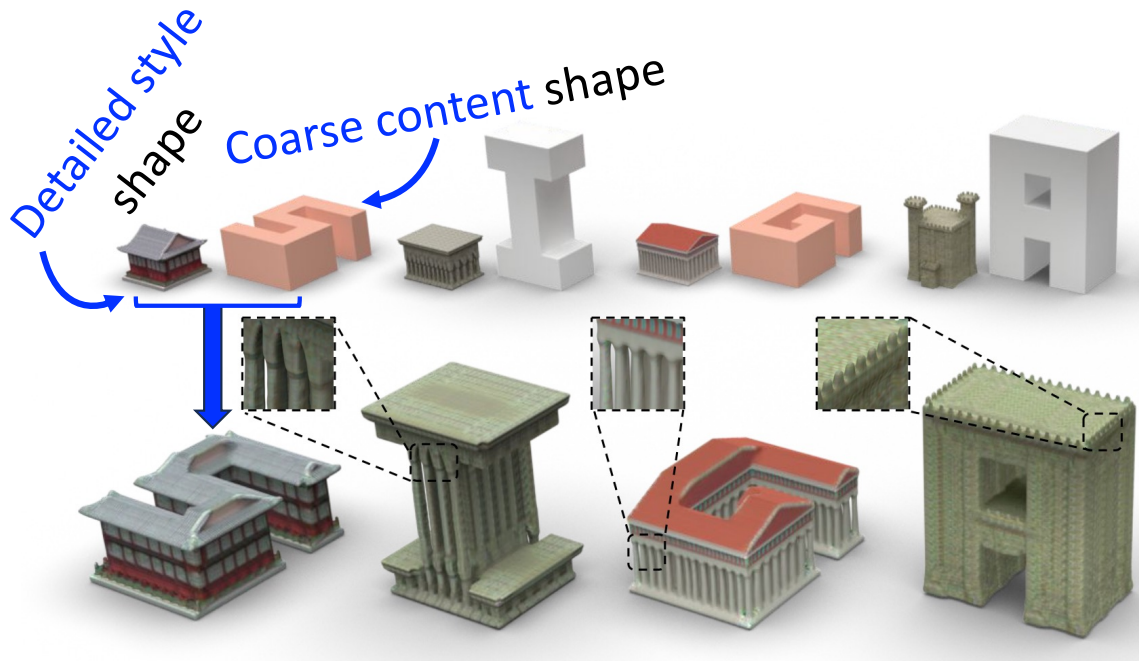
ARO-Net trained on **only the Fertility model** with rotation and scaling



Do not go BIG at all cost. Strive to **simplify** and “**smartize**” in terms of 3D training data and network architecture, etc.

# Addressing controllability challenge

- ❖ Give creators coarse-level structural control, then “detailize”



ShaDDR: geometry detailization and texture generation [Chen et al. SIGGRAPH Asia 2023]



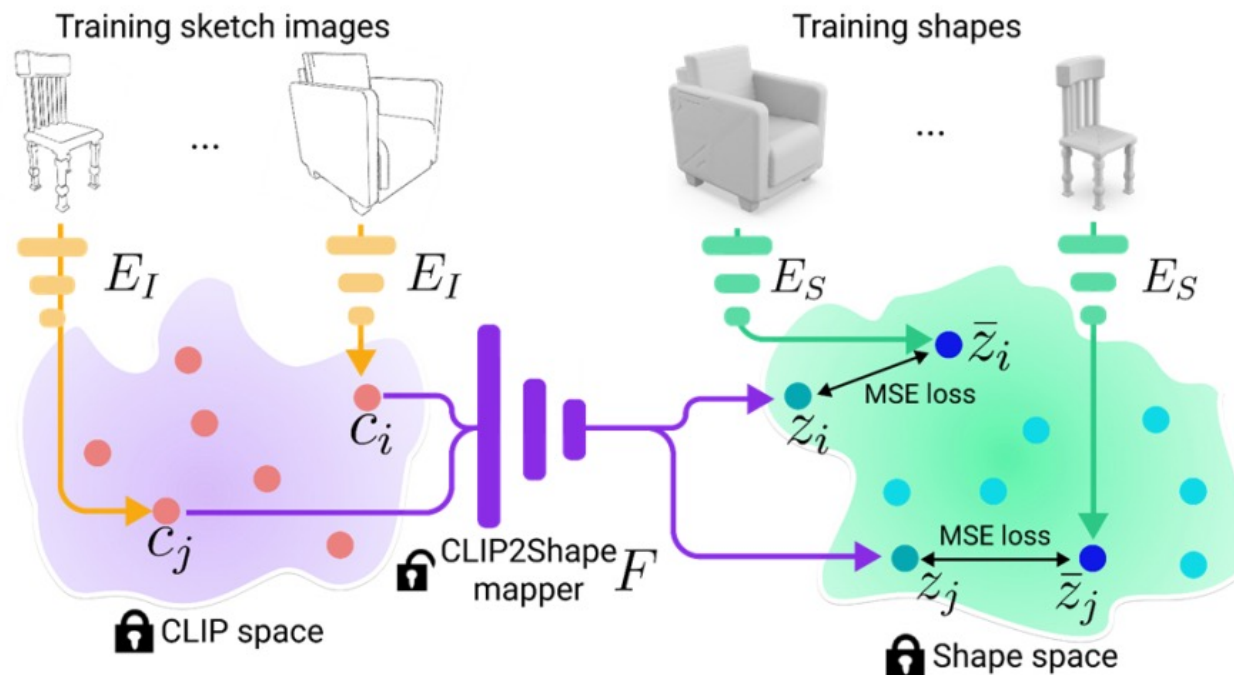


ShaDDR: geometry detailization and texture generation [Chen et al. SIGGRAPH Asia 2023]



# Explicitly enforcing text control

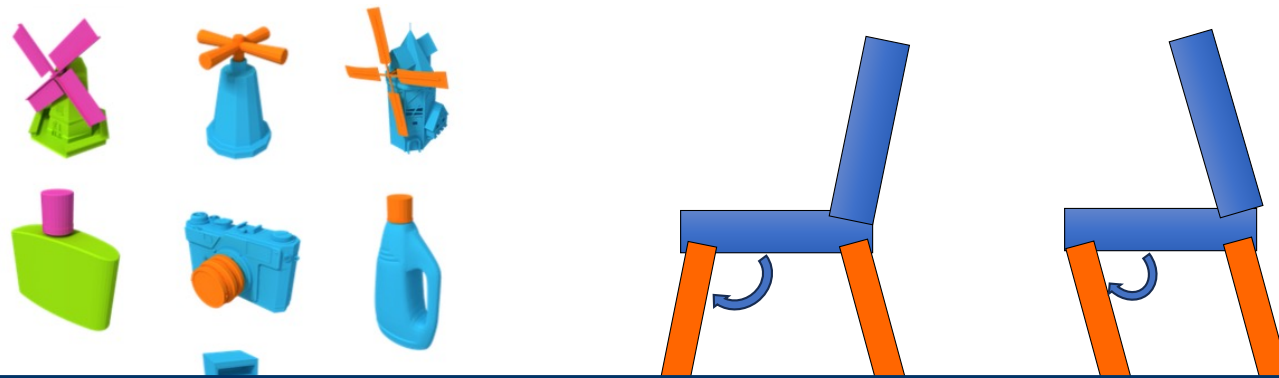
- ❖ Key: coupling CLIP and shape spaces with paired CLIP and shape codes



CLIPXplore  
[Lu et al. SIG Asia 2023]

# Addressing functionality challenge

- ❖ Functionality is mainly characterized by object **structures**, i.e., **object parts + part relations**



Learn to generate **structured 3D representations**, going beyond voxels, Gaussian splats, and moving into higher-level **primitives**

# From Minsky's "The Society of Mind" (1986)

## 12.5 THE FUNCTIONS OF STRUCTURES

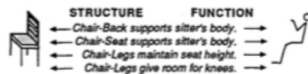
Many things that we regard as physical are actually psychological. To see why this is so, let's try to say what we mean by "chair." At first it seems enough to say:

"A chair is a thing with legs and a back and seat."

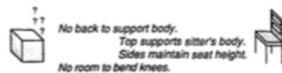
But when we look more carefully at what we recognize as chairs, we find that many of them do not fit this description because they don't divide into those separate parts. When all is done, there's little we can find in common to all chairs—except for their intended use.

"A chair is something you can sit upon."

But that, too, seems inadequate. It makes it seem as though a chair were as instantaneous as a wish. The solution is that we need to combine at least two different kinds of descriptions. On one side, we need structural descriptions for recognizing chairs when we see them. On the other side we need functional descriptions in order to know what we can do with chairs. We can capture more of what we mean by interweaving both ideas. But it's not enough merely to propose a vague association, because in order for it to have some use, we need more intimate details about how those chair parts actually help a person to sit. To catch the proper meaning, we need connections between parts of the chair structure and the requirements of the human body that those parts are supposed to serve. Our network needs details like these:



Without such knowledge, we might just crawl under the chair or try to wear it on our head. But with that knowledge we can do amazing things, like applying the concept of a chair to see how we could sit on a box, even though it has no legs or back!



Uniframes that include structures like this can be powerful. For example, such knowledge about relations between structure, comfort, and posture could be used to understand when a box could serve as a chair: that is, only when it is of suitable height for a person who does not require a backrest or room to bend the knees. To be sure, such clever reasoning requires special mental skills with which to redescribe or "reformulate" the descriptions of both box and chair so that they "match" despite their differences. Until we learn to make old descriptions fit new circumstances, our old knowledge can be applied only to the circumstances in which it was learned. And that would scarcely ever work, since circumstances never repeat themselves perfectly.

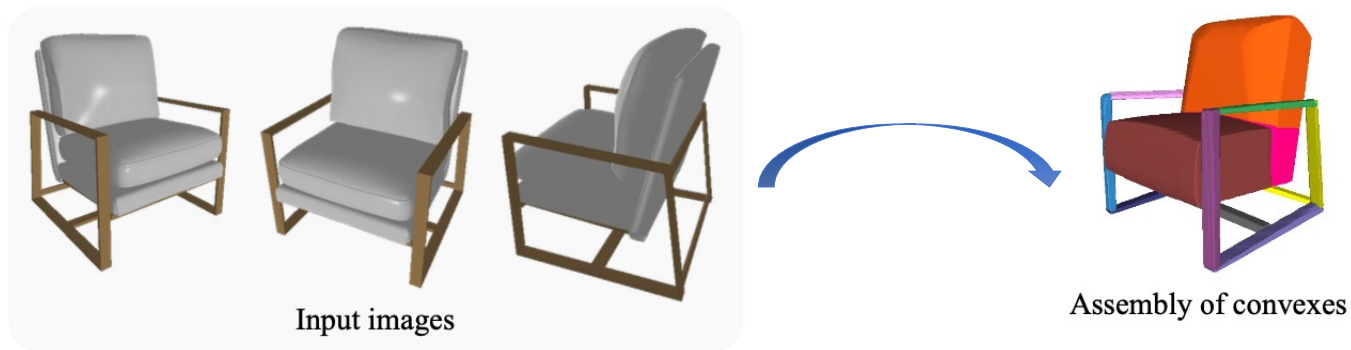
"There's little we can find in common to all chairs — except for their intended use."



"... we need to combine at least two different kinds of descriptions (of objects). On one side, we need **structural descriptions** for recognizing chairs when we see them. On the other side, we need **functional descriptions** in order to know what we can *do* with chairs."

# 3D CAD generation from sparse-view

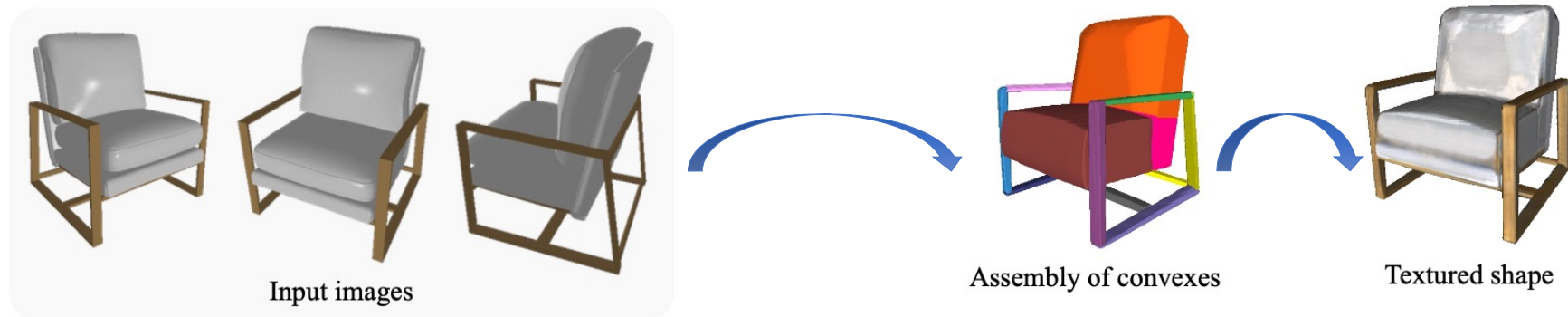
- ❖ Assembly of quadric primitives from sparse, wide-baseline views



[Yu et al. ECCV 2024]

# 3D CAD generation from sparse-view

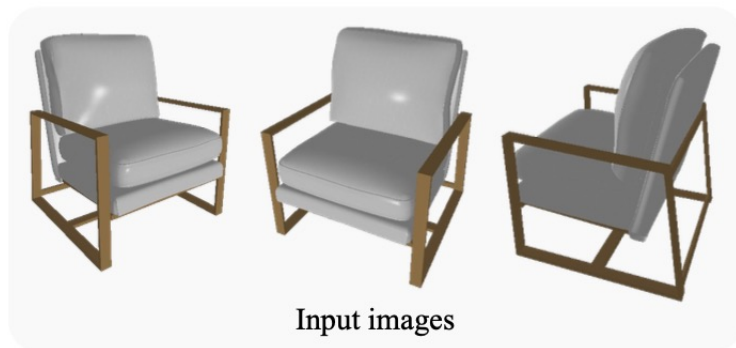
❖ Primitive assembly via **diff rendering**, w/o 3D supervision



[Yu et al. ECCV 2024]

# 3D CAD generation from s

❖ Generated primitive assembly **direct**



## CVPR 2023 by the Numbers

Selecting a category below changes the paper list on the right.

**SELECT** ↓ Top 10 overall by number of authors

		AUTHORS	PAPERS
1	3D from multi-view and sensors	1,090	246
2	Image and video synthesis and generation	889	185
3	Humans: Face, body, pose, gesture, movement	813	166
4	Transfer, meta, low-shot, continual, or long-tail learning	688	153
5	Recognition: Categorization, detection, retrieval	673	139
6	Vision, language, and reasoning	631	118
7	Low-level vision	553	126
8	Segmentation, grouping and shape analysis	524	113
9	Deep learning architectures and techniques	485	92
10	Multi-modal learning	450	89
11	3D from single images	431	91
12	Medical and biological vision, cell microscopy	420	53
13	Video: Action and event understanding	373	83
14	Autonomous driving	359	69
15	Self-supervised or unsupervised representation learning	349	71
16	Datasets and evaluation	344	54
17	Scene analysis and understanding	276	54
18	Adversarial attack and defense	274	61
19	Efficient and scalable vision	252	48
20	Computational imaging	226	53
21	Video: Low-level analysis, motion, and tracking	215	46
22	Vision applications and systems	171	35
23	Vision + graphics	155	32
24	Robotics	141	23
25	Transparency, fairness, accountability, privacy, ethics in vision	126	50

80%+ papers on “multi-view/sensor to 3D” target neural fields (NeRFs, 3DGS). More attention should on **multi-view to CAD!**

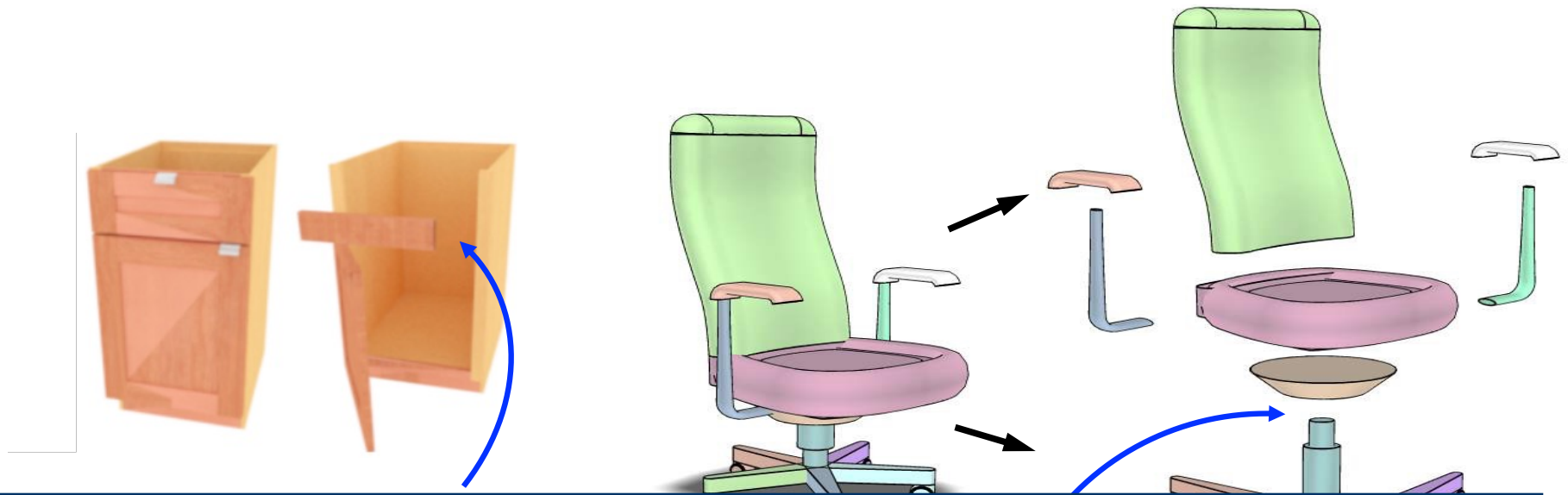
[Yu et al. ECCV 2022]

33	Optimization methods (other than deep learning)	46	12
34	Photogrammetry and remote sensing	38	8
35	Computer vision theory	33	5
36	Computer vision for social good	25	5



# Data challenge to functional reasoning

- ❖ Functioning objects often reveal **hidden/interior** structures



Dominant majority of digital 3D models in existence are missing interior structures for them to function — essentially “useless” 😊

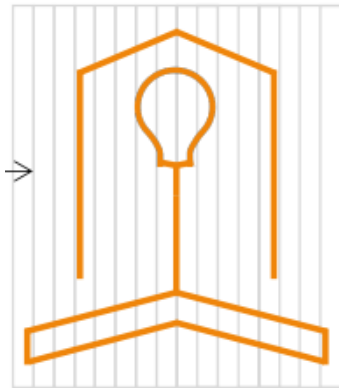
# Data challenge to functional reasoning



~2,300 models from PartNet-Mobility [Xiang et al. CVPR 2020]

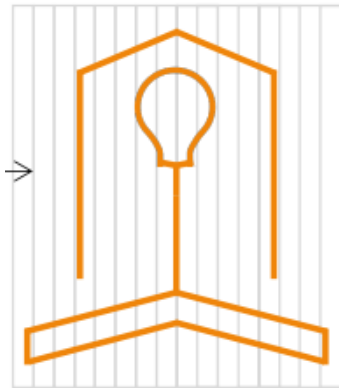
# Potential for generating interiors

- ❖ Key idea: use **slicing** to reveal object interiors



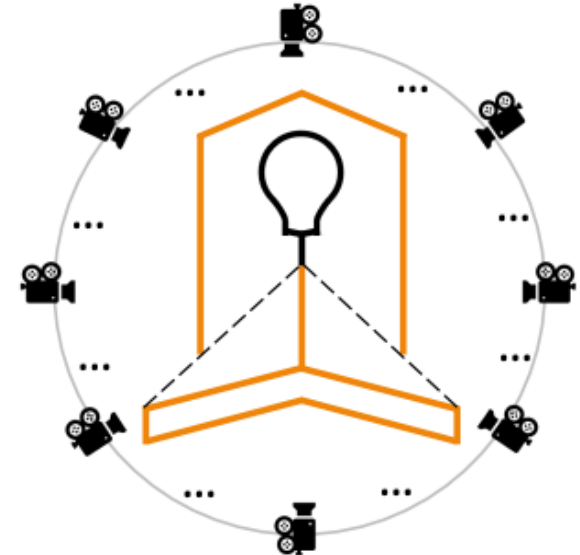
# Potential for generating interiors

❖ Key idea: use **slicing** to reveal object interiors (vs. multi-view)



Infinity many slices reveal everything!  
Also **sidesteps multi-view inconsistency**.

VS



Infinity many cameras still  
cannot reveal the bulb

# Slice3D

- ❖ Single image to 3D by first **generating multi-slice images**

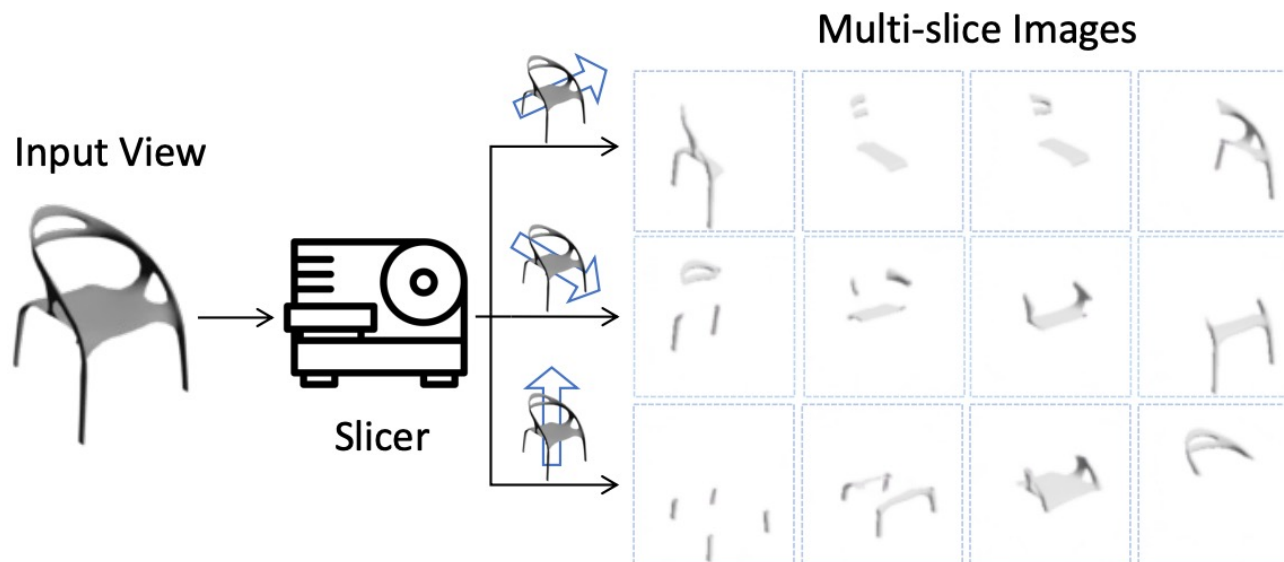
Input View



[Wang et al. CVPR 2014]

# Slice3D

- ❖ Single image to 3D by first **generating multi-slice images**

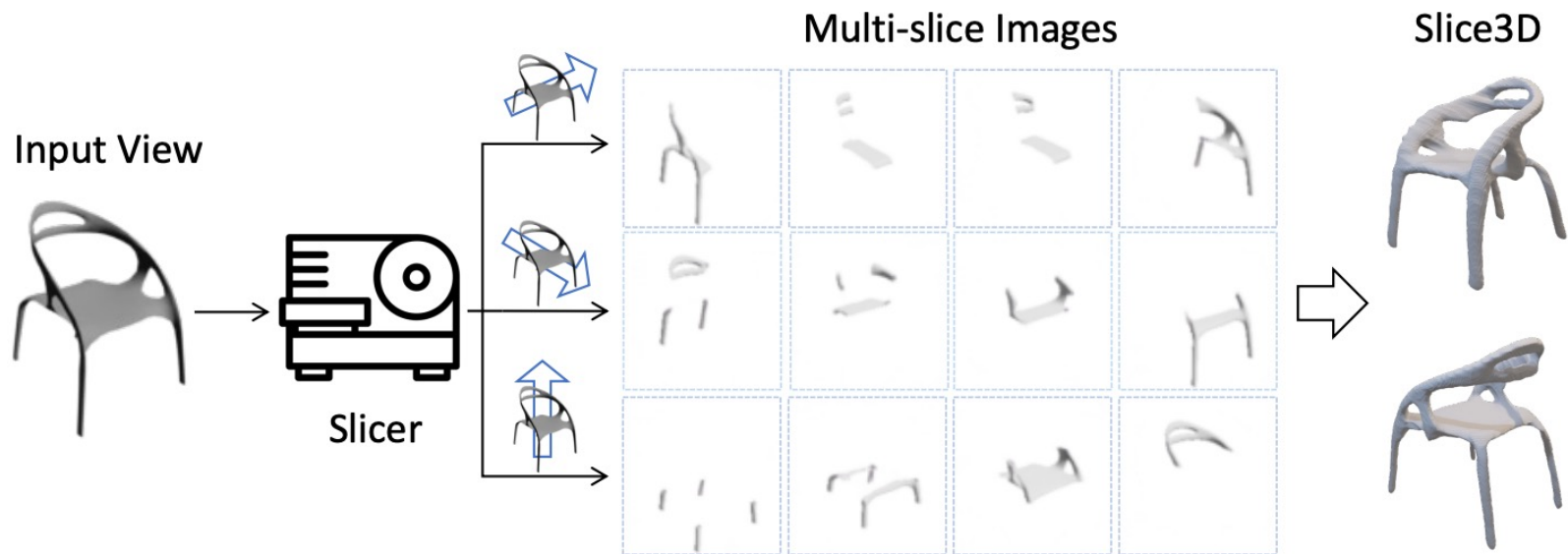


[Wang et al. CVPR 2014]



# Slice3D

- ❖ Single image to 3D by first **generating multi-slice images**



[Wang et al. CVPR 2014]

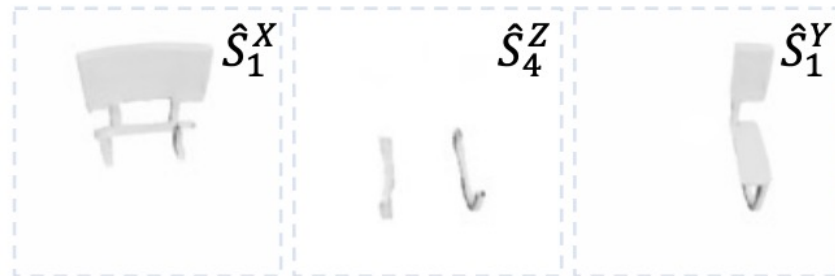
# Multi-slice generator via diffusion

- ❖ Diffusion-based slice generator produces **multiple plausible results**

Input image



Generated slices #1



Reconstructed mesh #1



[Wang et al. CVPR 2014]

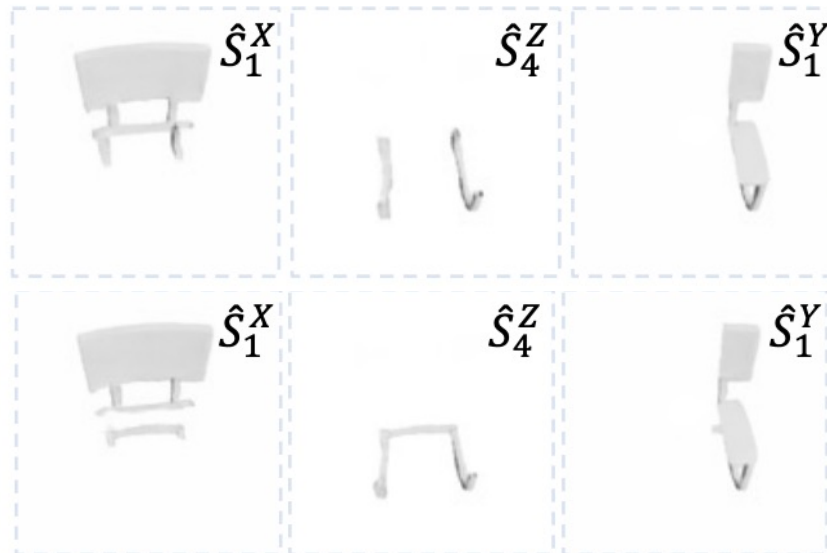
# Multi-slice generator via diffusion

- ❖ Diffusion-based slice generator produces **multiple plausible results**

Input image



Generated slices #1



Reconstructed mesh #1



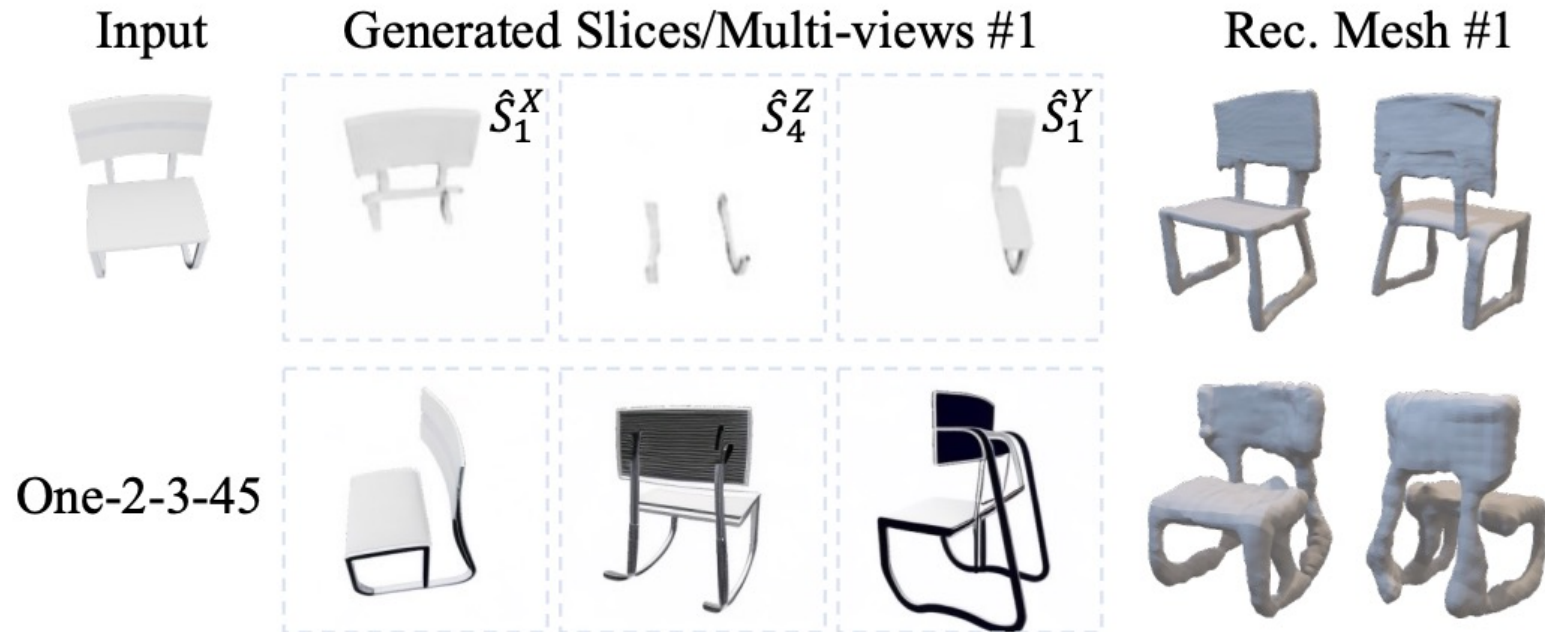
Reconstructed mesh #2



[Wang et al. CVPR 2014]

# Multi-slice vs. multi-view

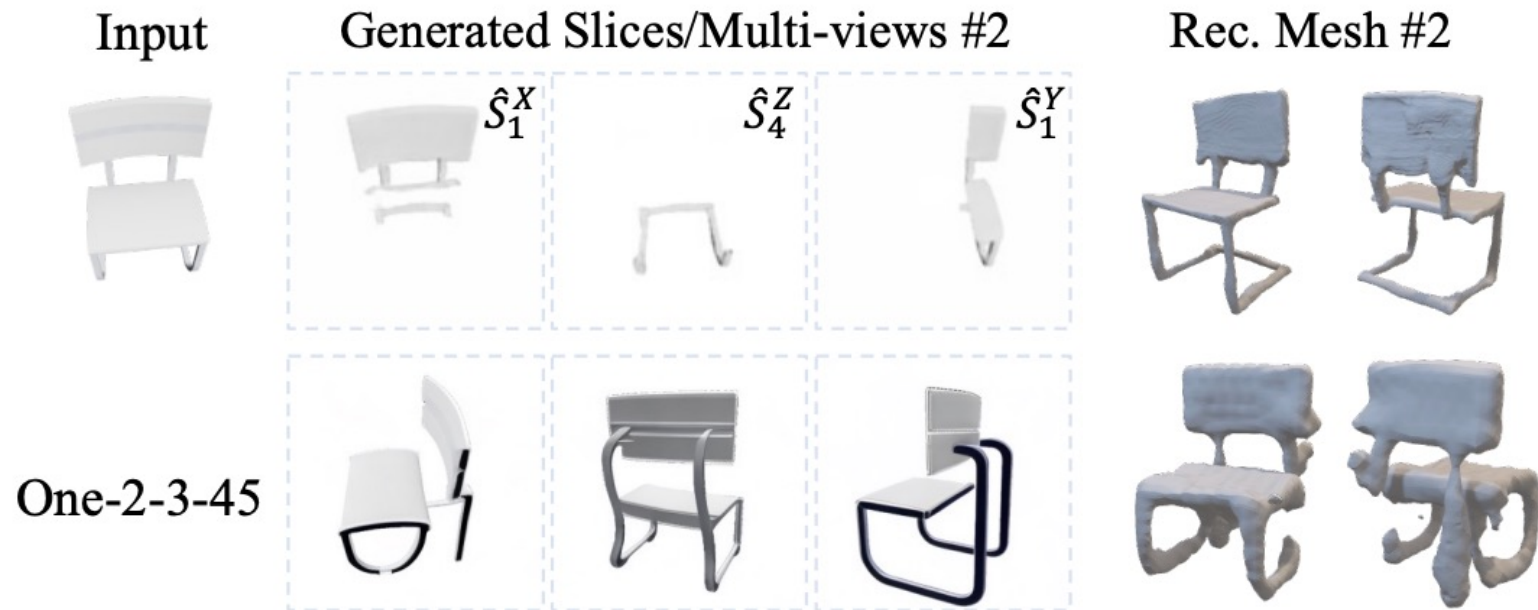
- ❖ Multi-view synthesis often suffers from **inconsistency problems**



[Wang et al. CVPR 2014]

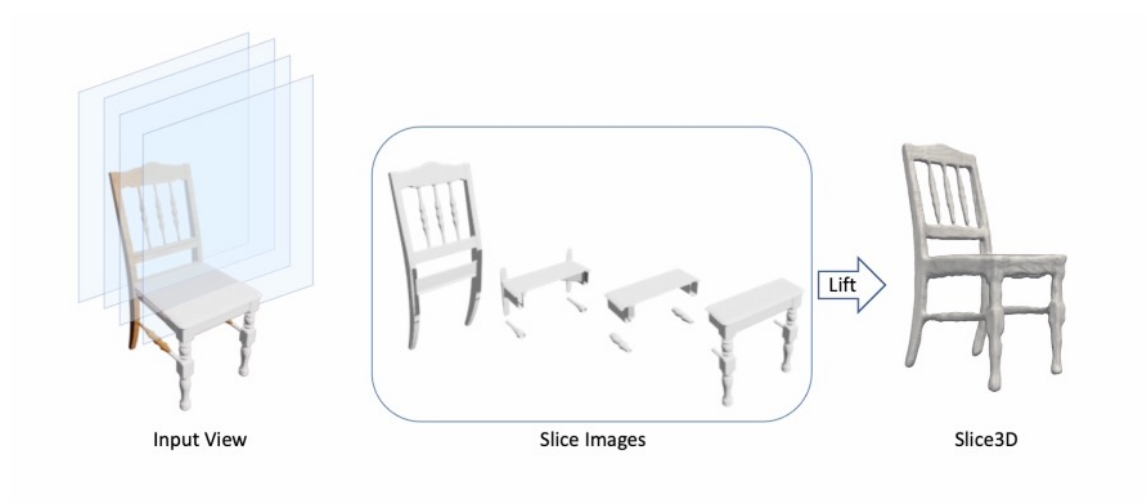
# Multi-slice vs. multi-view

- ❖ Multi-view synthesis often suffers from **inconsistency problems**



[Wang et al. CVPR 2014]

# Slice3D vs. SOTA



❖ No pre-trained LMs, e.g., Stable Diffusion



# Slice3D vs. SOTA



- ❖ No pre-trained LMs, e.g., Stable Diffusion
- ❖ Slice3D model trained using **5% of Objaverse 1.0** dataset on a **single Nvidia A40** GPU yields **inference time of ~20s**

# Slice3D vs. SOTA



- ❖ No pre-trained LMs, e.g., Stable Diffusion
- ❖ Slice3D model trained using 5% of Objaverse 1.0 dataset on a single Nvidia A40 GPU yields inference time of ~20s
- ❖ Example of “going small” with big gain from right presentation

# Potential for generating interiors

## ❖ Text-to-image-to-3D



# Potential for generating interiors

## ❖ Text-to-image-to-3D



EASI-TeX  
[Perla et al. SIGGRAPH 2024]

# Potential for generating interiors

❖ Text-to-image-to-3D with **generation of interiors**



# Active 3D reconstruction with robots



New trend in visual computing on “learning by interacting”, embodied AI, and increasing integration between vision, graphics, and robotics



# Addressing geometry challenge

- ❖ Ultimately, geometry is purported to **serve** intended **functions**



# Addressing geometry challenge

- ❖ Ultimately, geometry is purported to serve intended functions
- ❖ Both are strongly tied to motion, interaction, as well as physics



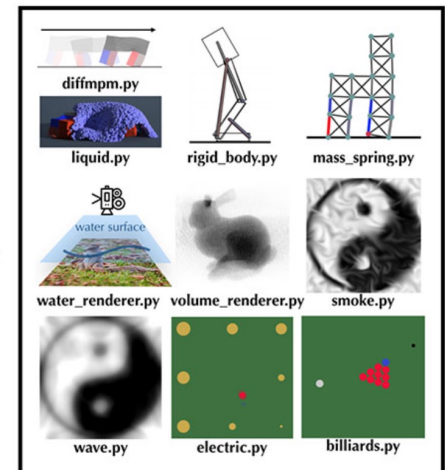
# Addressing geometry challenge

- ❖ Ultimately, geometry is purported to serve intended functions
- ❖ Both are strongly tied to motion, interaction, as well as physics
- ❖ Differentiable physics and physics-informed NNs (PINNs)

## DiffTaichi: Differentiable Programming for Physical Simulation

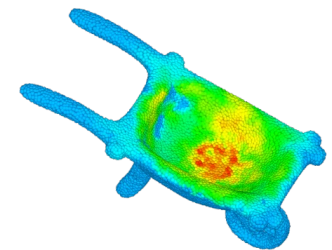
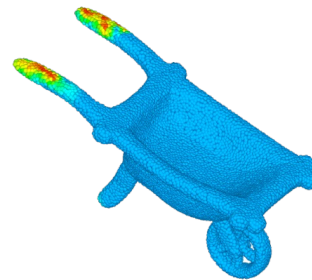
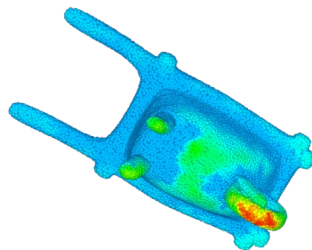
Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr,  
Jonathan Ragan-Kelley, Frédo Durand  
ICLR 2020

$$\frac{\partial \text{Yin-Yang}}{\partial \mathbf{X}}$$



# Addressing geometry challenge

- ❖ Ultimately, geometry is purported to serve intended functions
- ❖ Both are strongly tied to motion, interaction, as well as physics
- ❖ Differentiable physics and physics-informed NNs (PINNs)
- ❖ “Differentiable function analysis” still elusive, but an “attention layer” for functional interactions is worth exploring



[Hu et al. SIGGRAPH 2016]

# Summary

😊 LFM's will continue to play a key role in 3D GenAI due to their **zero-shot** and **generalization** capabilities, **but**

# Summary

😊 LFM will continue to play a key role in 3D GenAI due to their zero-shot and generalization capabilities, but

- 👉 Text-to-X not the best paradigm: ambiguity and lack of control
- 👉 **Smart** use of LMs as losses (e.g., CLIP) or through prompting
- 👉 Consider **geometric/physical/functional properties** when designing prompts, chain-of-thought reasoning, etc.
- 👉 Keep **humans in the loop** with high-level priors and controls



# Summary

😞 LFM's are **too costly**: do not go BIG at all costs, instead,

# Summary

- ☹️ LFM's are too costly: do not go BIG at all costs, instead,
- 👉 Data-centric AI with use of small and “smartized” 3D data
- 👉 Continuing pursuit of the “best” 3D representation
  - ❖ Shifting from voxels/points for rendering to modeling primitives, with emphases on structures, motions, and functions
  - ❖ Maybe the best representation is ultimately a hybrid one

# Summary

AI research

Nov 22, 2023

**Microsoft's Orca 2 can beat LLMs 5-10 times its size thanks to a unique training method**



Check out **SLM = Small Language Models**: more efficient, lower-cost (60-200M parameters vs. 175B for ChatGPT), and more customizable

# Back to “Geometry Challenge”

*Geometry is a branch of mathematics concerned with the properties of **space** such as the distance, shape, size, and relative position of figures.*

Wikipedia

# SPACE: Spatial Perception And Cognition Evaluation

[Submitted on 9 Oct 2024]

## Does Spatial Cognition Emerge in Frontier Models?

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, Vladlen Koltun

Not yet. We present SPACE, a benchmark that systematically evaluates spatial cognition in frontier models. Our benchmark builds on decades of research in cognitive science. It evaluates large-scale mapping abilities that are brought to bear when an organism traverses physical environments, smaller-scale reasoning about object shapes and layouts, and cognitive infrastructure such as spatial attention and memory. For many tasks, we instantiate parallel presentations via text and images, allowing us to benchmark both large language models and large multimodal models. Results suggest that contemporary frontier models fall short of the spatial intelligence of animals, performing near chance level on a number of classic tests of animal cognition.

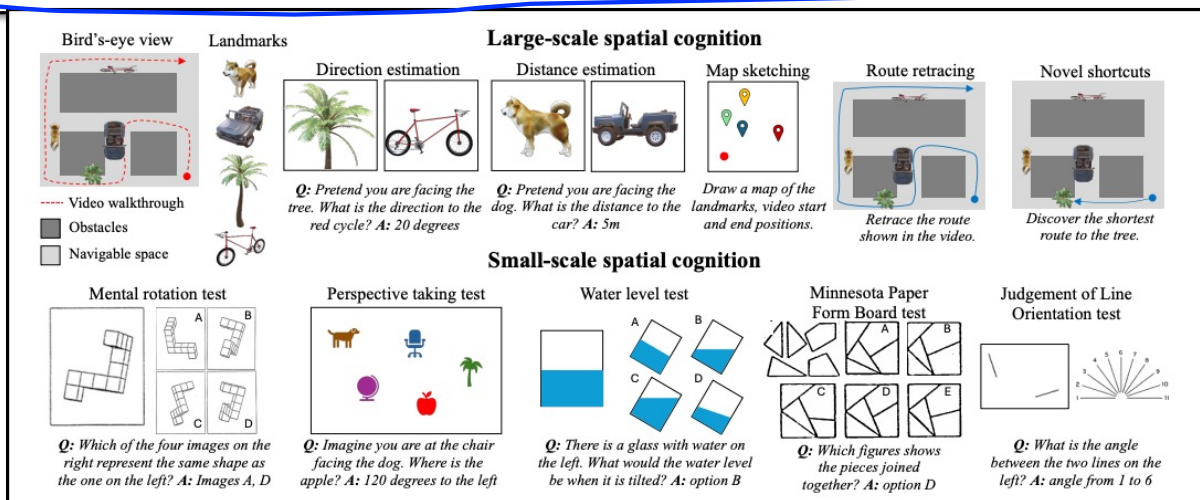
# SPACE: Spatial Perception And Cognition Evaluation

[Submitted on 9 Oct 2024]

## Does Spatial Cognition Emerge in Frontier Models?

Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, Vladlen Koltun

Not yet. We present SPACE, a benchmark that systematically evaluates spatial cognition in frontier models. Our benchmark builds on decades of research in cognitive science. It evaluates large-scale mapping abilities that are brought to bear when an organism traverses physical environments, smaller-scale reasoning about object shapes and layouts, and cognitive infrastructure such as spatial attention and memory. For many tasks, we instantiate parallel presentations via text and images, allowing us to benchmark both large language models and large multimodal models. Results suggest that contemporary frontier models fall short of the spatial intelligence of animals, performing near chance level on a number of classic tests of animal cognition.





# Next wave: spatial and physical AI



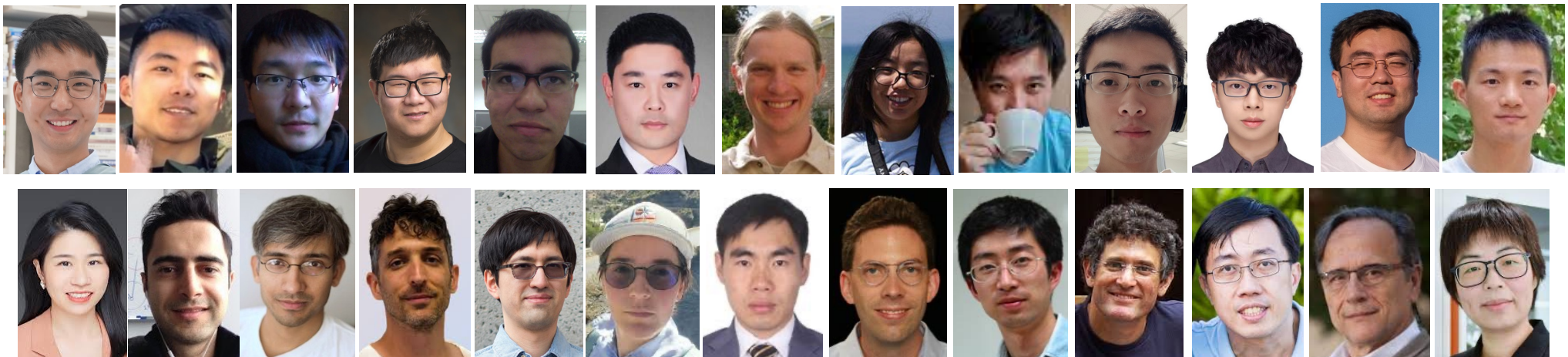
- ☹️ Contemporary LFM's far from having **spatial intelligence**
- 😊 To do the **right geometry/function/physics** is key to 3D GenAI

# Main publications covered

- [Chen et al. 2023] (ShaDDR) Qimin Chen, Zhiqin Chen, Hang Zhou, and Hao Zhang, “ShaDDR: Interactive Example-Based Geometry and Texture Generation via 3D Shape Detailization and Differentiable Rendering,” *SIGGRAPH Asia* 2023.
- [Chen et al. 2023] (Décollage) Qimin Chen, Zhiqin Chen, Vova Kim, Noam Aigerman, Siddhartha Chaudhuri, and Hao Zhang, “Décollage: 3D Detailization by Controllable, Localized, and Learned Geometry Enhancement,” *ECCV* 2024.
- [Hu et al. 2023] (CLIPXplore) Jingyu Hu, Kai-Hei Hui, Zhengzhe Liu, Hao Zhang, and Chi-Wing Fu, “CLIPXplore: Coupled CLIP and Shape Spaces for 3D Shape Exploration,” *SIGGRAPH Asia* 2023.
- [Hu et al. 2016] Ruizhen Hu, Oliver van Kaick, Bojian Wu, Hui Huang, Ariel Shamir, and Hao Zhang, “Learning How Objects Function via Co-Analysis of Interactions,” *SIGGRAPH* 2016.
- [Ma et al. 2018] Rui Ma, Akshay Gadi Patil, Matthew Fisher, Manyi Li, Soren Pirk, Binh-Son Hua, Sai-Kit Yeung, Xin Tong, Leonidas Guibas, and Hao Zhang, “Language-Driven Synthesis of 3D Scenes from Scene Datasets,” *SIGGRAPH Asia* 2023.
- [Wang et al. 2023] (ARO-Net) Yizhi Wang, Zeyu Huang, Ariel Shamir, Hui Huang, Hao Zhang, and Ruizhen Hu, “ARO-Net: Learning Implicit Fields from Anchored Radial Observations,” *CVPR* 2023.
- [Wang et al. 2023] (Slice3D) Yizhi Wang, Wallace Lira, Wenqi Wang, Ali Mahdavi-Amiri, and Hao Zhang, “Slice3D: Multi-Slice, Occlusion-Revealing, Single View 3D Reconstruction,” *CVPR* 2024.
- [Yan et al. 2023] Zihao Yan, Fubao Su, Mingyang Wang, Ruizhen Hu, Hao Zhang, and Hui Huang, “Interaction-Driven Active 3D Reconstruction with Object Interiors,” *SIGGRAPH Asia* 2023.
- [Yu et al. 2024] (DPA-Net) Fenggen Yu, Yiming Qian, Xu Zhang, Francisca Gil-Ureta, Brian Jackson, Eric Bennett, and Hao Zhang, “DPA-Net: Structured 3D Abstraction from Sparse Views via Differentiable Primitive Assembly” *ECCV* 2024.

# Acknowledgment

❖ Students, postdocs, and collaborators on covered works



*Adobe, Amazon, Carleton University, Jilin University, Reichman University, Chinese University of Hong Kong, Shandong University, Shenzhen University, Stanford University, and University of Montreal*