Lecture 10:

Generative AI for Content Creation (Part III)

Visual Computing Systems
Stanford CS348K, Spring 2024

Reminder: key aspect in the design of any system

Choosing the "right" representations for the job

- **Good representations are productive to use:**
  - Embody a "preferred" way of thinking about a problem

- **Good representations enable the system to provide useful services:**
  - Validating/providing certain guarantees (correctness, resource bounds, conversion of quantities, type checking)
  - Performance optimizations (parallelization, vectorization, use of specialized hardware)
  - Implementations of common, difficult-to-implement functionality (complex array indexing code, texture mapping in 3D graphics, auto-differentiation, etc.)
  - Execute a complex edit that the user has in their head

Here: choosing the right representation is choosing controls that are most useful to an editing task

- What is the type of control that aligns with the users thought process / mental model of editing?
  - Text is often an ambiguous, imprecise, or flat out inefficient way to describe visual intent

- Examples:
  - Users want to control spatial composition
    - "Dog on the left" vs. dragging a layer to the right location
  - Users want to "block out" an idea, and have the diffusion model "fill in the details", "correct proportions", "harmonize the image"
  - Users want to express intent via an example: "I want it to look LIKE THIS!"

Discussion:

Propose a type of edit that you would like to make to images

How does the user “think” about what they are trying to change (are they worried about details, composition, a particular “axis” of change (e.g, adjust smile but not eyes)

How could to generate supervision to train a model to support this type of control?

Generating other forms of media: Videos, 3D meshes, animation, etc…

Scarsity of data

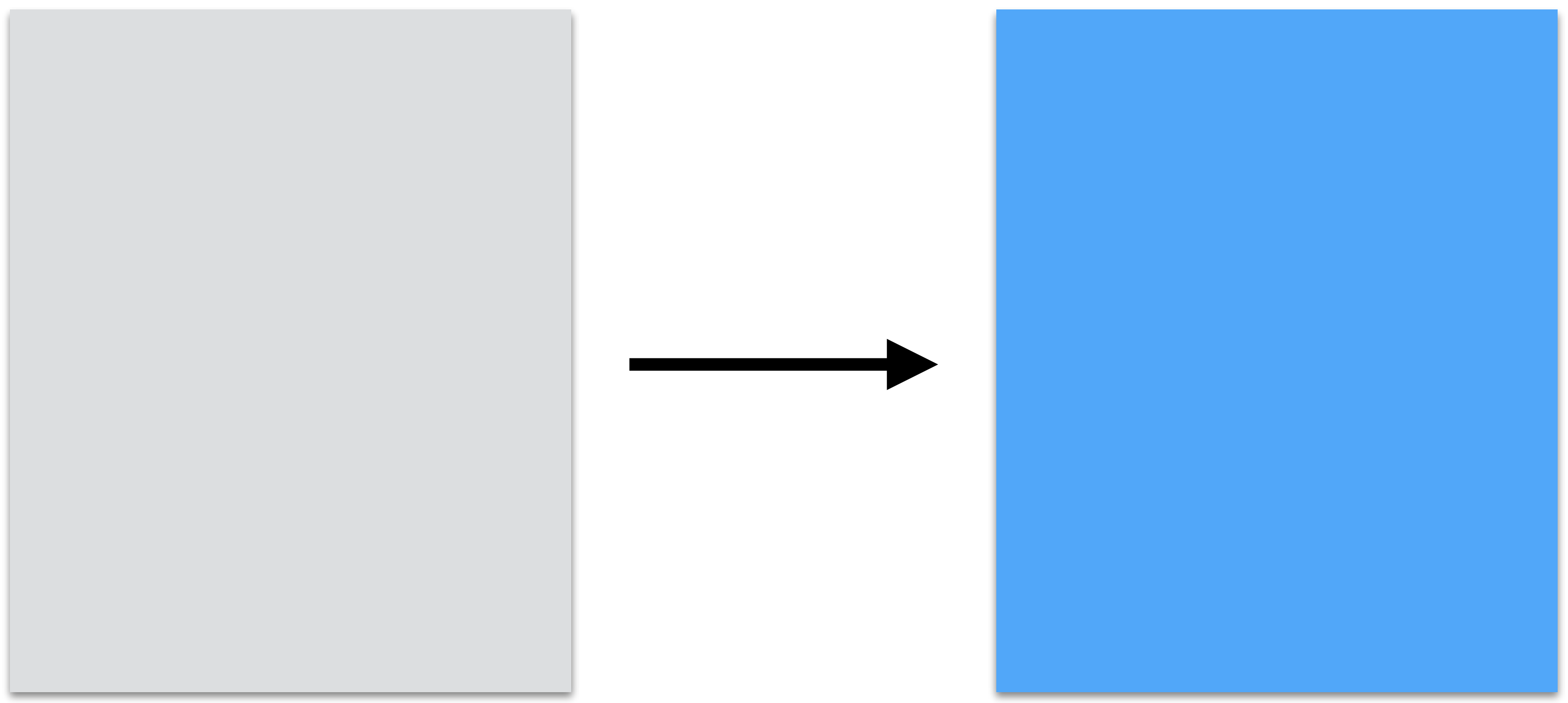
- Recall that text-to-image generation models were trained on billions of image-text pairs
- But datasets of paired video, 3D models, animation, etc. do not exist at this scale
- So most techniques for generating other forms of media start with models trained on images and ("lift") them to other forms of media

Video diffusion examples

One example: text to video

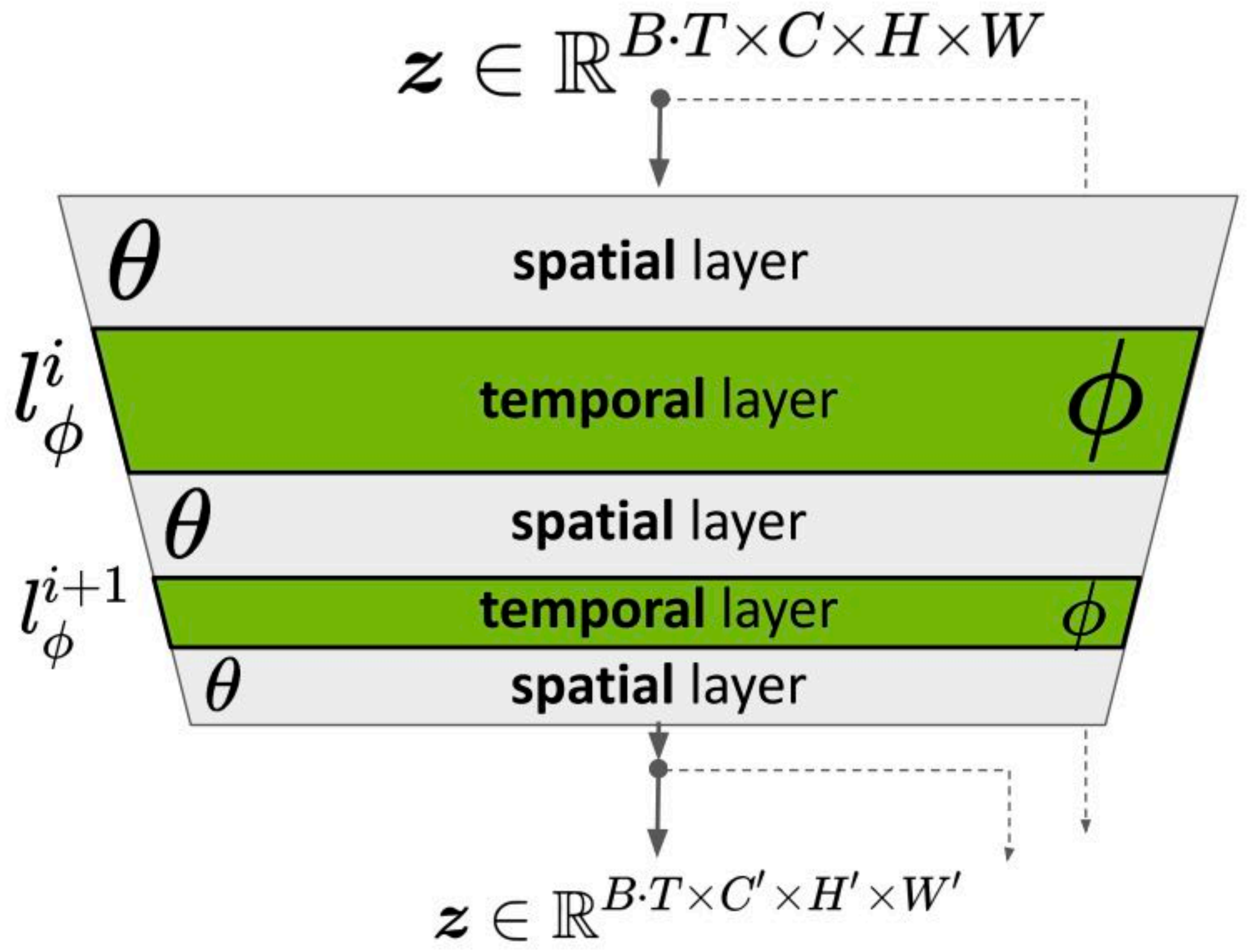
- Spatial blocks can be pretrained on large image datasets (visual visual representations)
- Temporal blocks trained on [smaller] video datasets

TxWxH TxWxH



“Spatial block” (T independent frames of size WxH)

“Temporal block” Mixes information across T frames

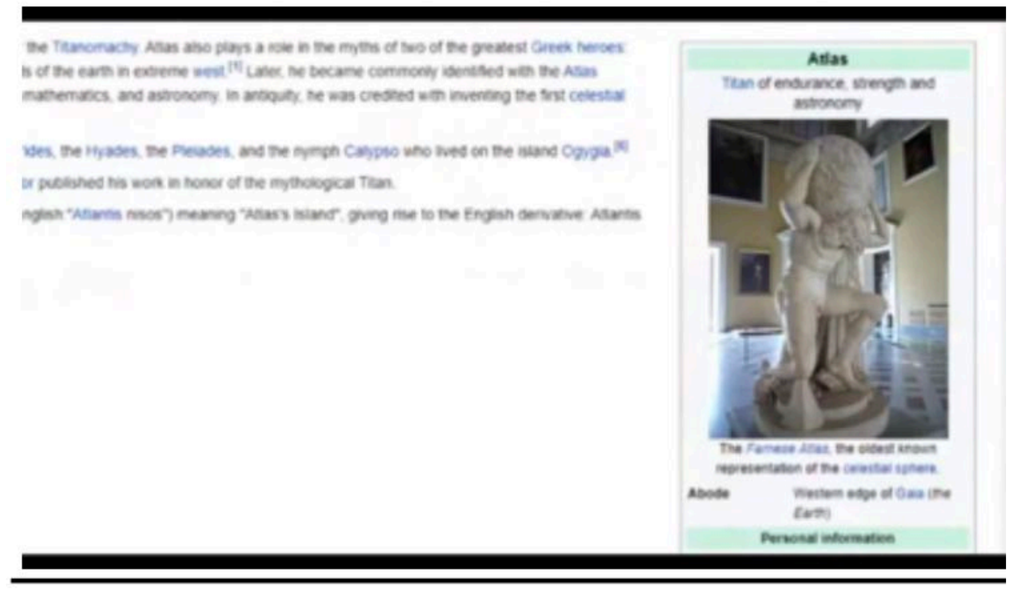


Example from [Blattman et al. 2023]

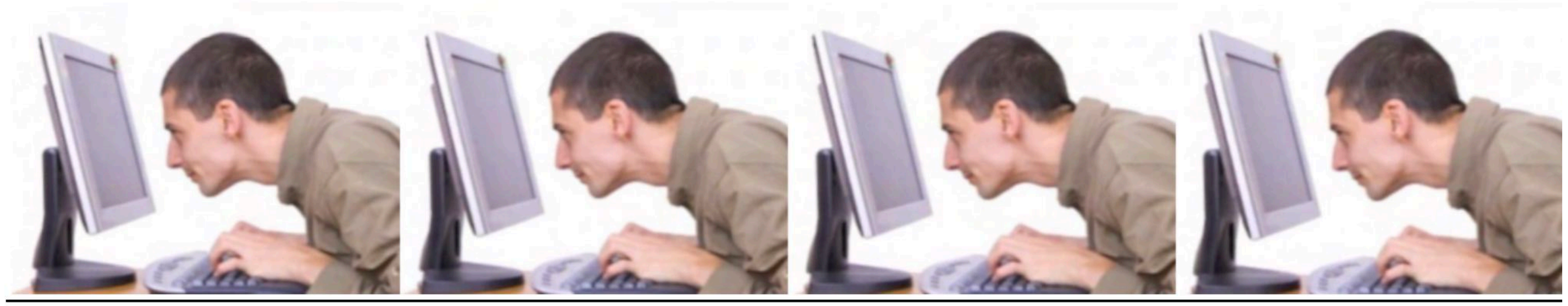
Dirty secret of modern ML: good dataset engineering

- **Given that video datasets are smaller, notable benefit to careful curation of video data training sets**

Internet videos have cuts / crossfades



Or screenshots/captions, where the screen is filled with text



Long periods of still frames

Also: use Visual Language models (VLM) or image captioning models to auto-caption the videos

Image credit: Stability.AI

Controlling video generation

- New controls emerging rapidly
  - Object control
  - Camera control

3D object diffusion

Story so far…

- **Given paired C1 -> X2, train a diffusion model…**

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathbf{z}_i, \quad i = 0, 1, \cdots, T$$

("score function")

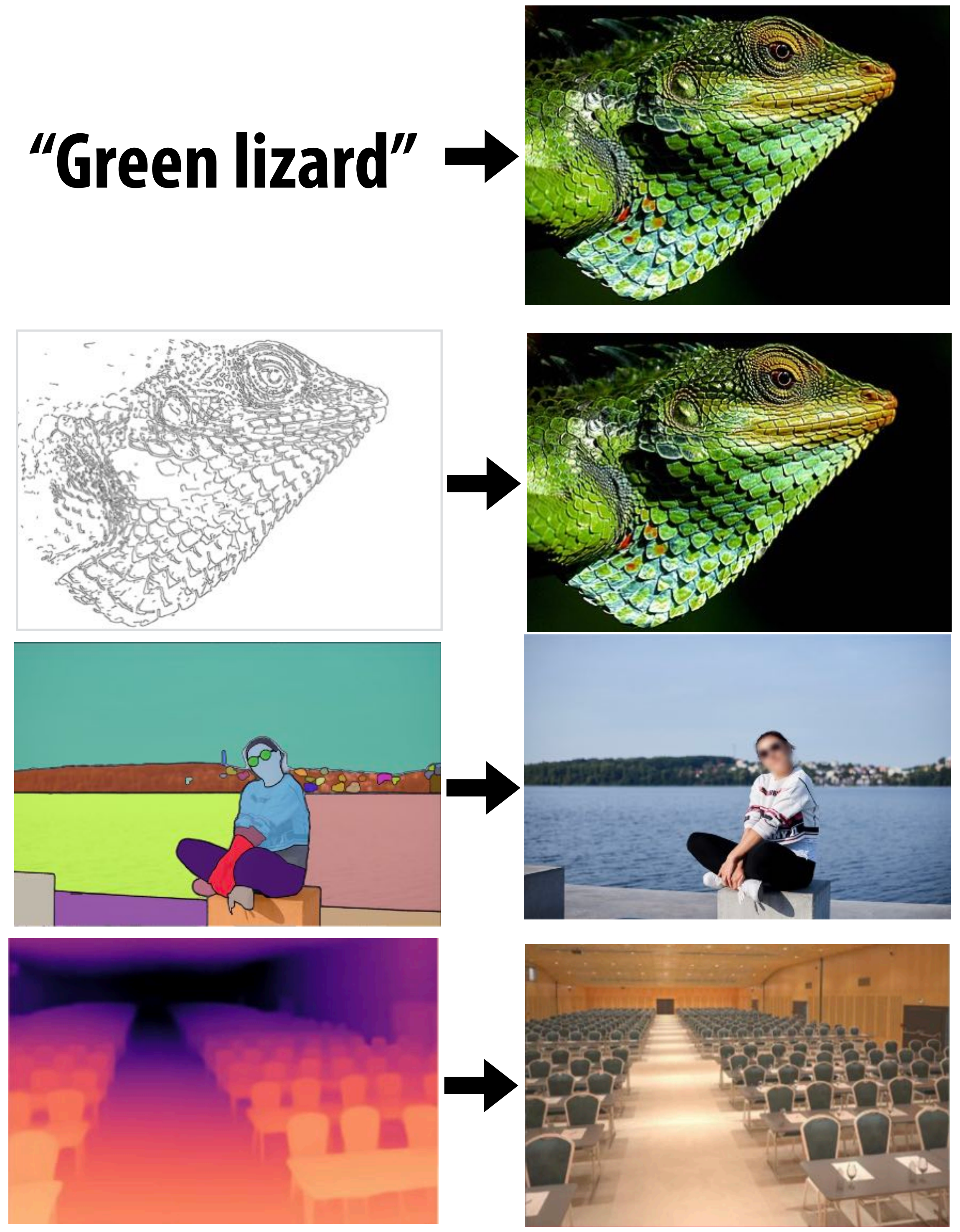
$$\nabla_{\mathbf{x}} \log p(\mathbf{x} \mid \mathbf{y}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}) + \nabla_{\mathbf{x}} \log p(\mathbf{y} \mid \mathbf{x})$$

(Unguided score function)

Modify image x so that image is more likely
[to come from the BILLION IMAGE training set]

(Prompt guidance)

Modify image x to make the CONDITIONING a
more likely related to the image



But now we want conditioning → 3D model

- But we don't have the datasets to learn the distribution of 3D models
- But we do know:
  - What the distribution of real images (what a realistic image is)
  - How to turn a 3D model into an image (rendering)

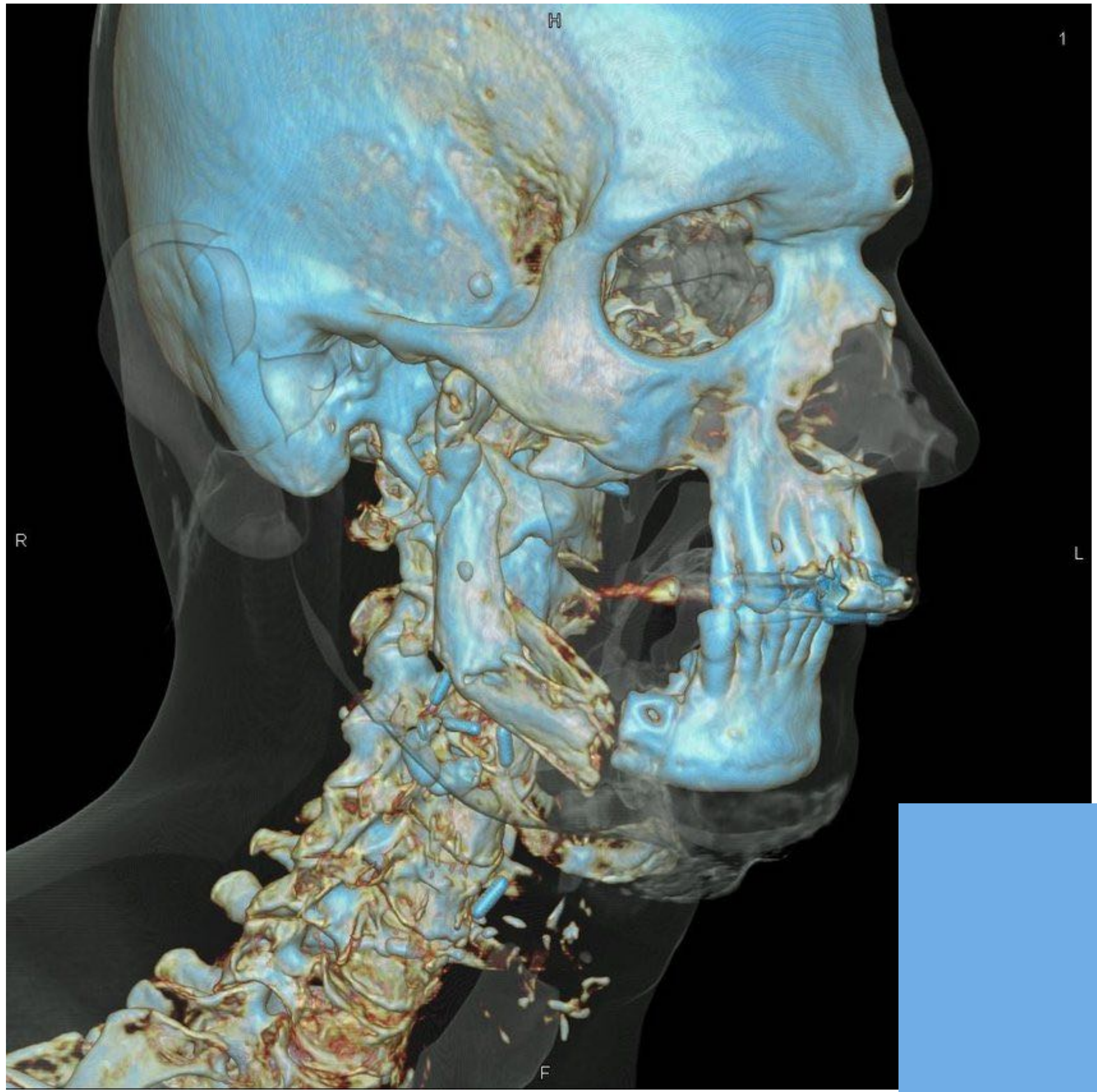
Let's represent 3D objects as volumes

Volume density and "color" at all points in space.

$$\sigma(\mathrm{p})$$

$$c(\mathrm{p}, \omega) = c(x, y, z, \phi, \theta)$$

The reflectance off surface at point p in direction ω



Aside: rendering volumes

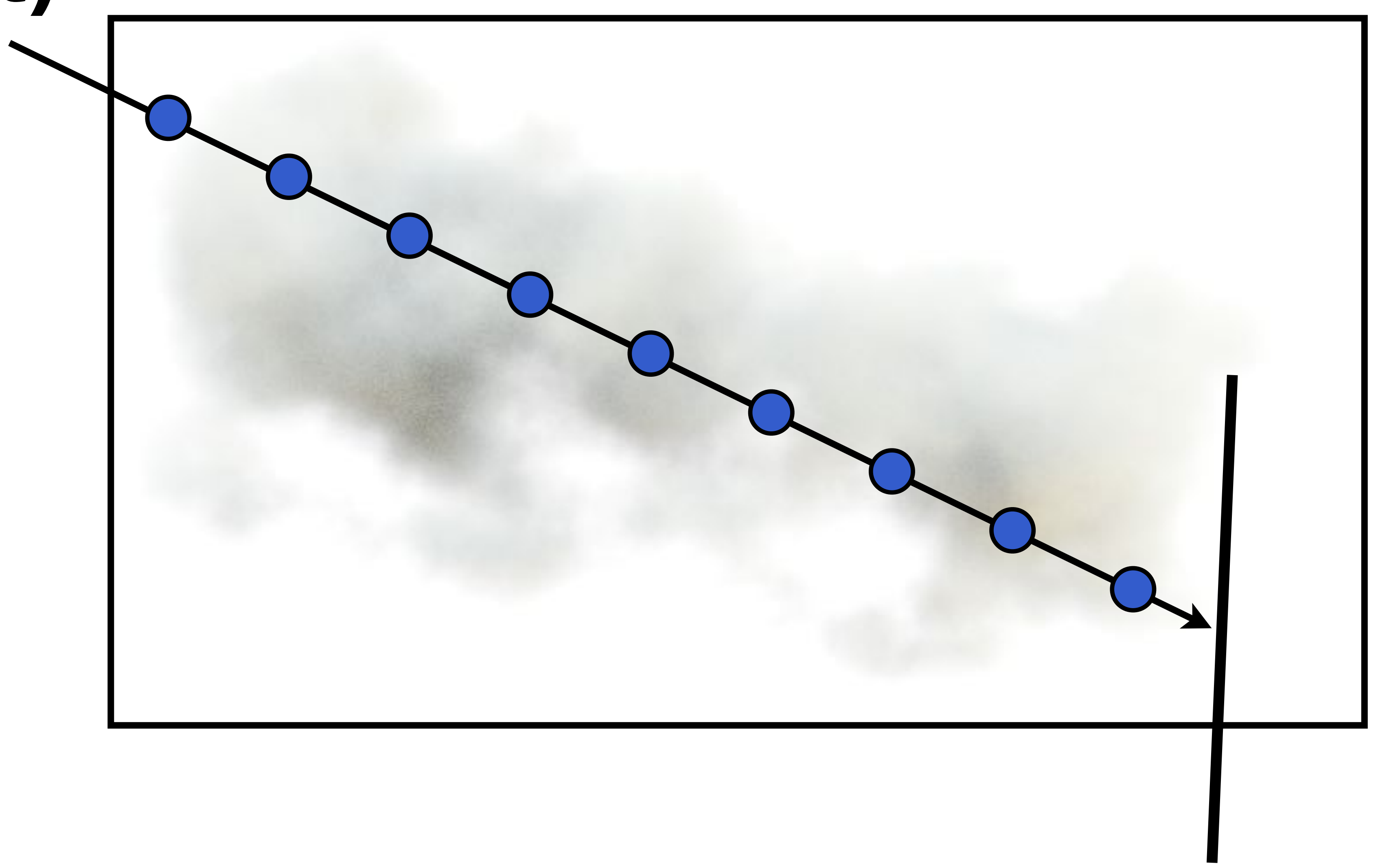
Given "camera ray" from point o in direction w….

$$\mathbf{r}(t) = \mathbf{o} + t\omega$$

And volume with density and directional radiance.

$$\sigma(\mathrm{p})$$
$$c(\mathrm{p}, \omega)$$

Volume density and color at all points in space.



Step through the volume to compute radiance along the ray.

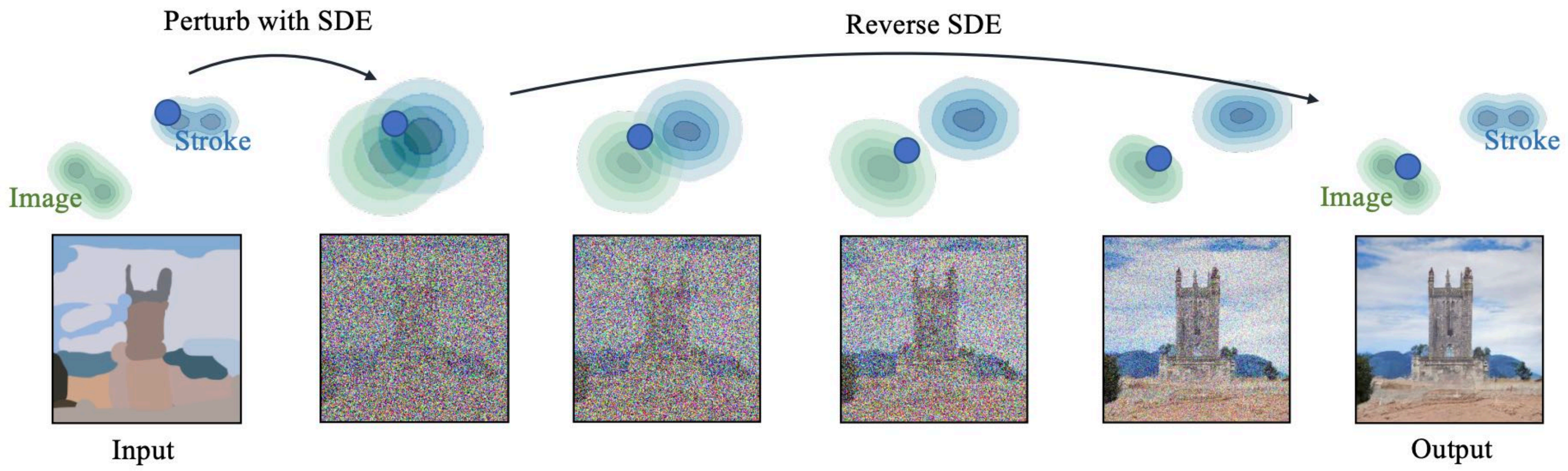
$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

Distilling a 3D generation model

- The takeaway: we can make an image from a 3D representation (e.g., a volume) using a differentiable rendering function

$$\mathrm{x} = R(\theta)$$

- And, with a image diffusion model, we can push that image closer to the distribution of real images



1. Start with a guide image (a target)
2. Add "small" amount of noise
3. Iteratively denoise to produce sample from target image distribution

Distilling a 3D generation model

- The takeaway: we can make an image from a 3D representation (e.g., a volume) using a differentiable rendering function

$$\mathrm{x} = R(\theta)$$

- And, with a text-conditioned image diffusion model, we can push image closer to the distribution of real images associated with a given prompt. That diffusion denoting step produces…

$$\Delta \mathrm{x}$$

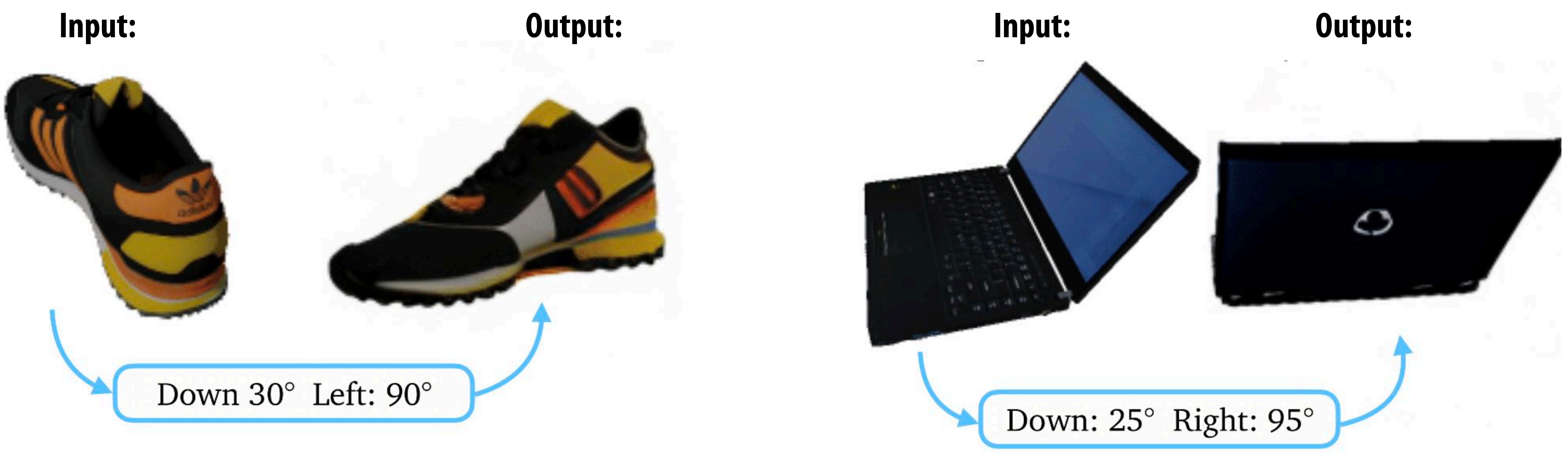
- Now, given $\nabla R(\theta)$ (recall R() was differentiable), optimize the parameters θ of the 3D representation to produce…

$$\mathrm{x} + \Delta \mathrm{x}$$

- In other words, we've converted the score function of a text-conditioned image diffusion model into a training procedure for a text-conditioned 3D diffusion model

Image-conditioned 3D diffusion

- Now let's say we want to condition 3D generation based on an image, not text:
- How about a simpler image editing problem: given a reference image X, and camera change parameters (rotation, translation), produce a novel view of the object in the image



Take a (pretty big) dataset of objects, fine tune image diffusion model on pairs

Rendered from viewpoint 1

Rendered from viewpoint 2

Input (conditioning)

Output

(X,R,T)

Now let's apply the same distillation trick

- We can now train an image conditioned 3D generation model using a similar process as described before in lecture

Animation diffusion

Text-conditioned animation generation

"A person walks forward, bends down to pick something up off the ground."

Audio-conditioned dance generation

Tseng et al CVPR 2023