Pre-class meet n' greet topics for your table:

If you were training a text-to-video generation model and had access to all of Youtube, what videos would you choose to train on?

What video would you not want to train on?

Lecture 8:

Generating Training Set for Visual Foundation Models

Visual Computing Systems

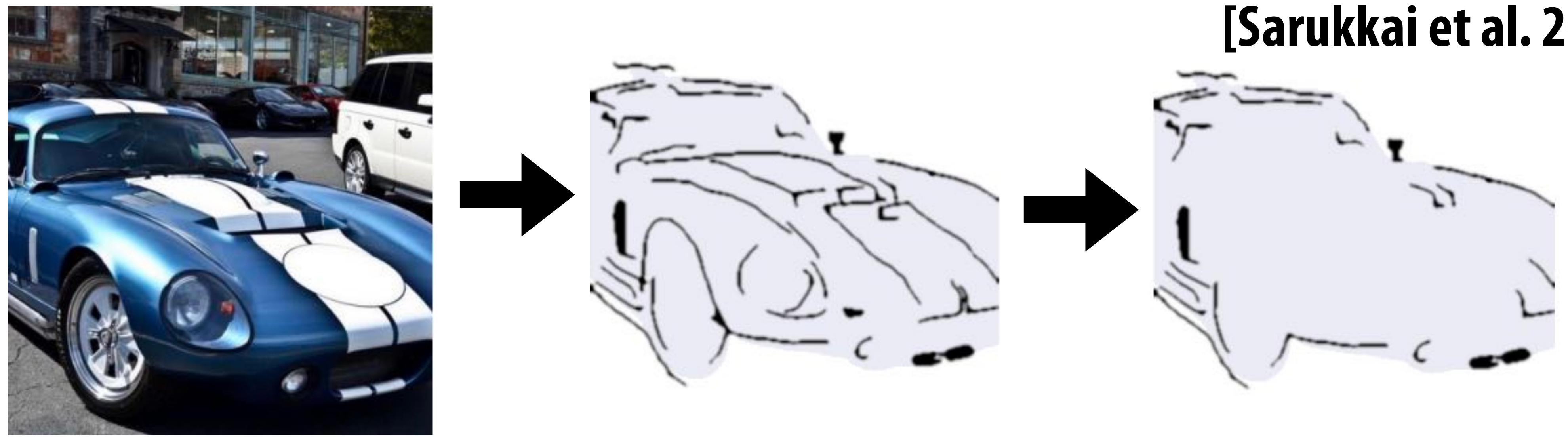
Stanford CS348K, Spring 2025

One topic from recent classes

- If you can generate paired training data, you can train a conditioned generation model
  - Text-to-X (Where X is image/video/3D model)
  - Image-to-X
  - Depth-to-image
  - Line sketch-to-image
  - Human pose-to-image

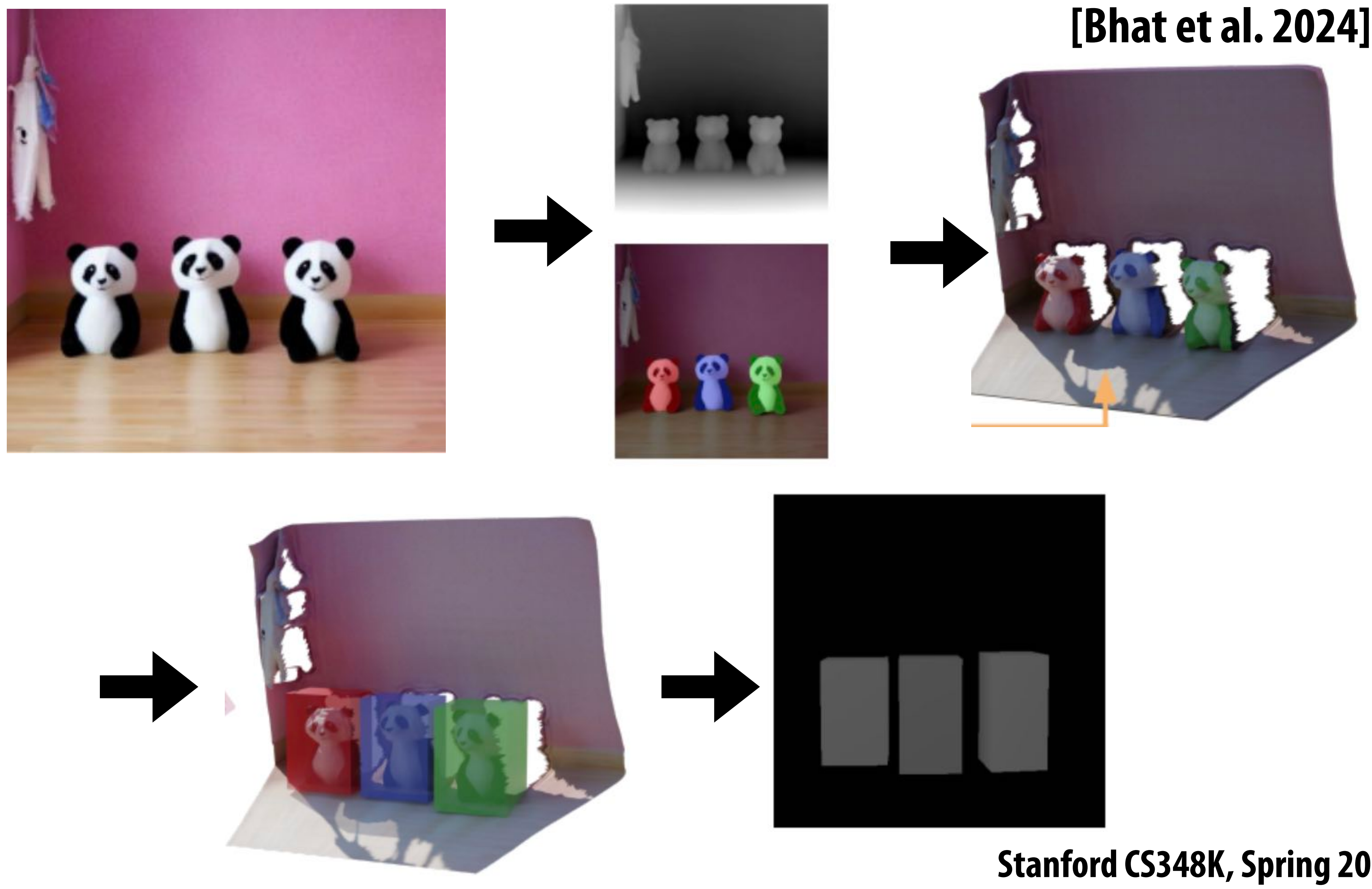
- For interesting conditions, the paired condition is generated using automatic methods ("pseudo annotation")
  - Impractical to annotate a sufficiently large dataset by hand
  - Usually even crowdsourcing is impractical

Stoke control: must estimate plausible strokes from image: detect edges, detect objects, retain edges near silhouettes



[Sarukkai et al. 2024]

Cuboid control: detect objects, estimate depth, estimate 3D boxes, render 2D projection of 3D boxes



[Bhat et al. 2024]

You are training a video generation model at Google.
You have access to all of Youtube's "free for use" content.

You are interested in text-to-video conditioning.
And maybe even sketch an object-movement path to video conditioning.

How would you go about selecting and annotating your data?

Training data selection (for video)

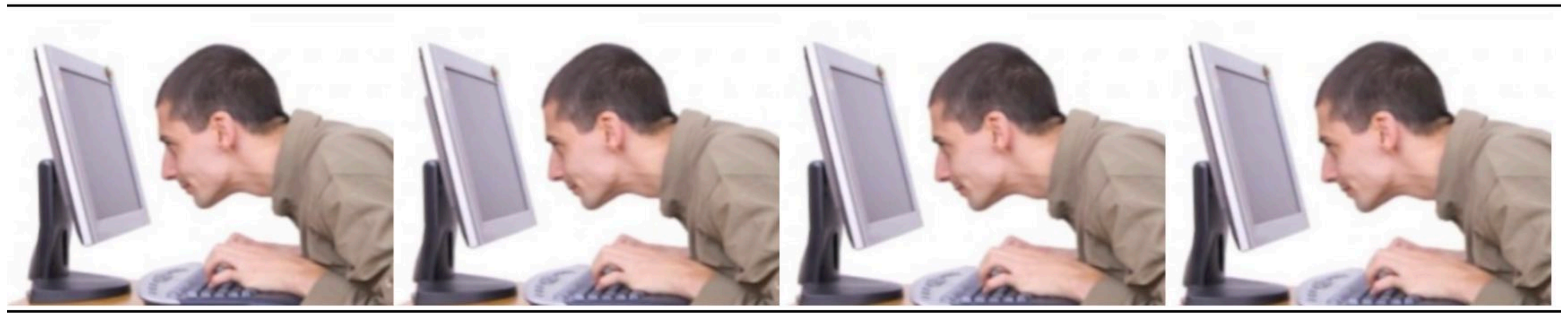
What content might we want to filter from training set? (Not train on)

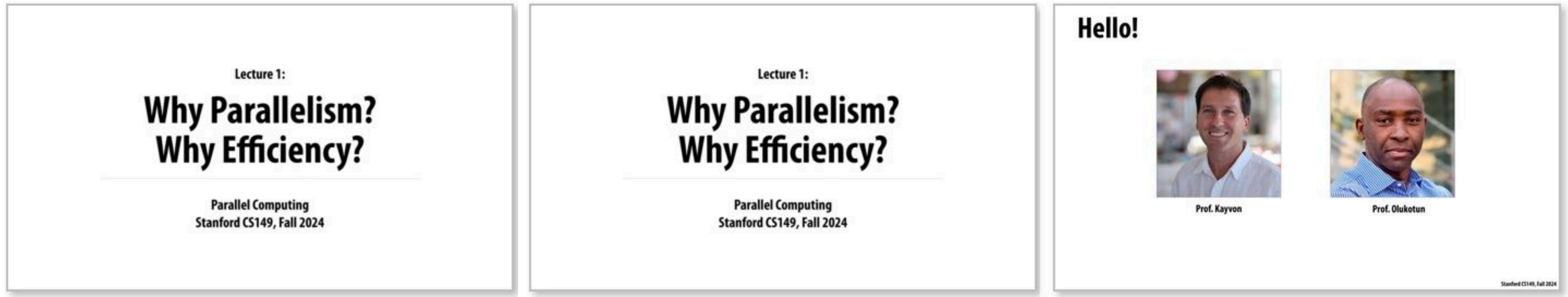
Inappropriate content

Duplicate content

Copyrighted content, (even if it’s marked as free use… could be duplicate of copyrighted content)

Static image videos (e.g., Kayvon’s lecture videos on Youtube)





Lecture 1:

Why Parallelism?
Why Efficiency?

Parallel Computing
Stanford CS149, Fall 2024

Lecture 1:

Why Parallelism?
Why Efficiency?

Parallel Computing
Stanford CS149, Fall 2024

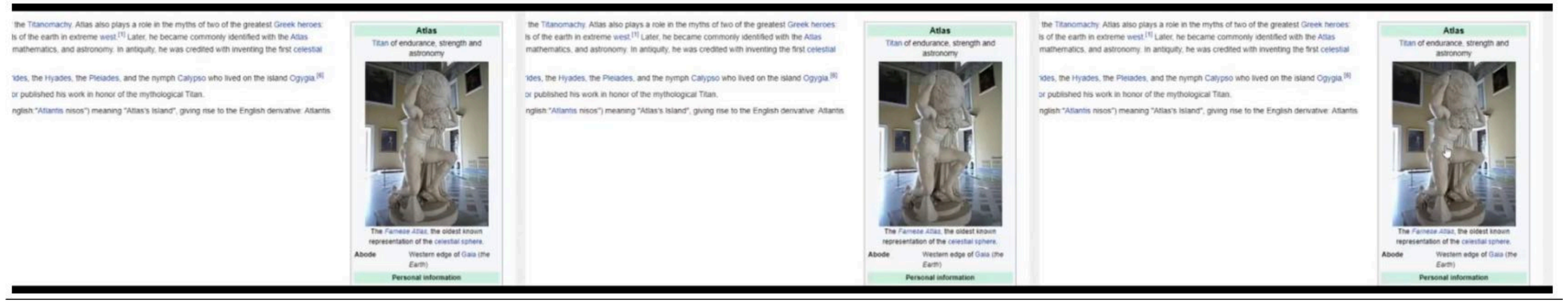
Hello!

Prof. Kayvon

Prof. Olukotun

Stanford CS149, Fall 2024

Videos with large amounts of text (often presentations/slides)



the Titanomachy. Atlas also plays a role in the myths of two of the greatest Greek heroes:
ls of the earth in extreme west.[1] Later, he became commonly identified with the Atlas
mathematics, and astronomy. In antiquity, he was credited with inventing the first celestial

ides, the Hyades, the Pleiades, and the nymph Calypso who lived on the island Ogygia.[6]
or published his work in honor of the mythological Titan.
nglish:"Atlantis nisos") meaning "Atlas's Island", giving rise to the English derivative: Atlantis

Atlas

Titan of endurance, strength and astronomy

The Farnese Atlas, the oldest known representation of the celestial sphere.

Abode | Western edge of Gaia (the Earth)

Personal information

the Titanomachy. Atlas also plays a role in the myths of two of the greatest Greek heroes:
ls of the earth in extreme west.[1] Later, he became commonly identified with the Atlas
mathematics, and astronomy. In antiquity, he was credited with inventing the first celestial

ides, the Hyades, the Pleiades, and the nymph Calypso who lived on the island Ogygia.[6]
or published his work in honor of the mythological Titan.
nglish:"Atlantis nisos") meaning "Atlas's Island", giving rise to the English derivative: Atlantis

Atlas

Titan of endurance, strength and astronomy

The Farnese Atlas, the oldest known representation of the celestial sphere.

Abode | Western edge of Gaia (the Earth)

Personal information

the Titanomachy. Atlas also plays a role in the myths of two of the greatest Greek heroes:
ls of the earth in extreme west.[1] Later, he became commonly identified with the Atlas
mathematics, and astronomy. In antiquity, he was credited with inventing the first celestial

ides, the Hyades, the Pleiades, and the nymph Calypso who lived on the island Ogygia.[6]
or published his work in honor of the mythological Titan.
nglish:"Atlantis nisos") meaning "Atlas's Island", giving rise to the English derivative: Atlantis

Atlas

Titan of endurance, strength and astronomy

The Farnese Atlas, the oldest known representation of the celestial sphere.

Abode | Western edge of Gaia (the Earth)

Personal information

Videos with rapid cinematic cuts

Videos that have content that current generative AI models are known to struggle width

[Adopted from Blattman et al. 2023]

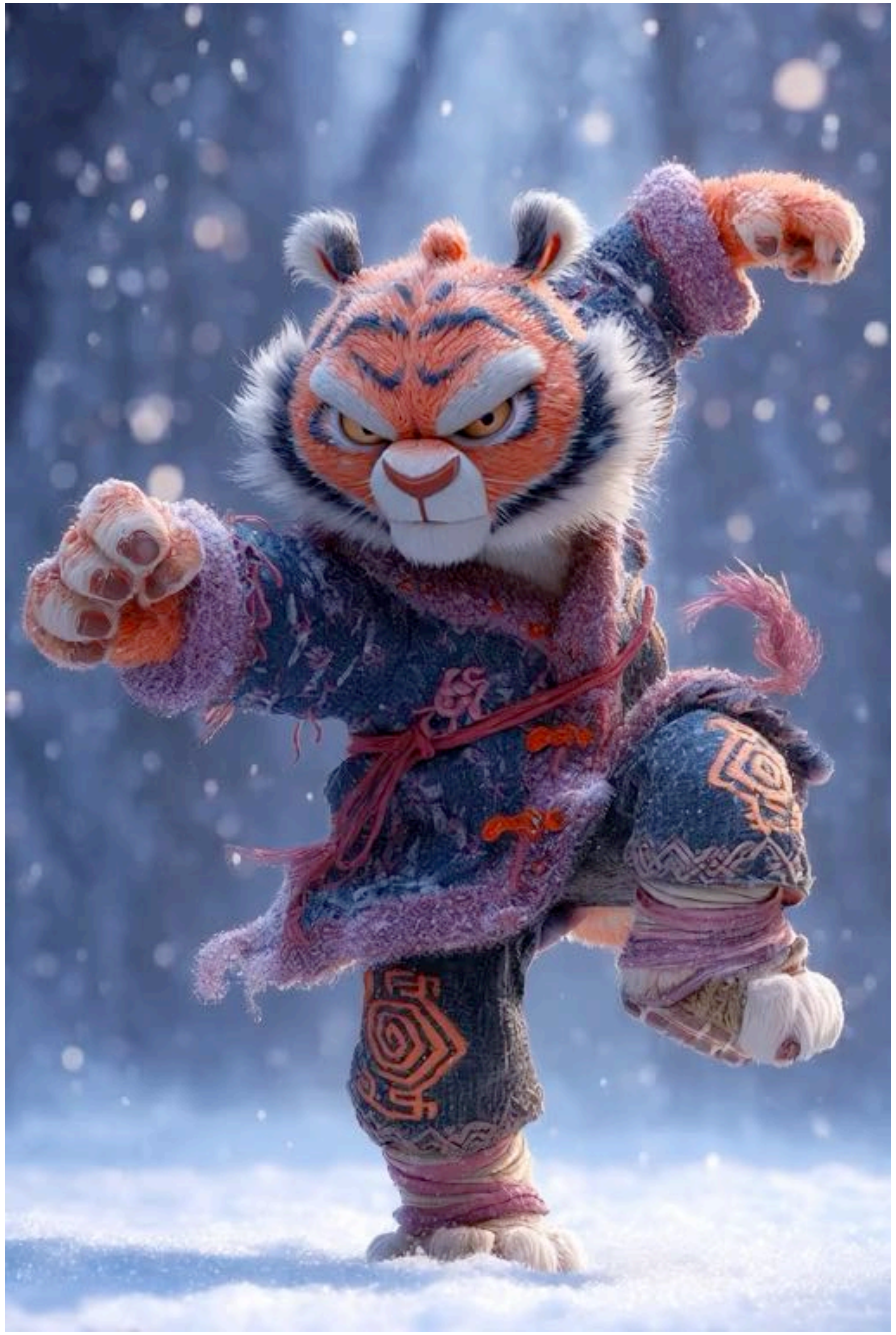
Training data annotation (for video)

What type of annotations might we want?

For text conditioning: text description of each video:

- But are properties of a “good” description?

What prompt would you use if you wanted to create this video?



[Generated by FramePack: Zhang et al. 2025]

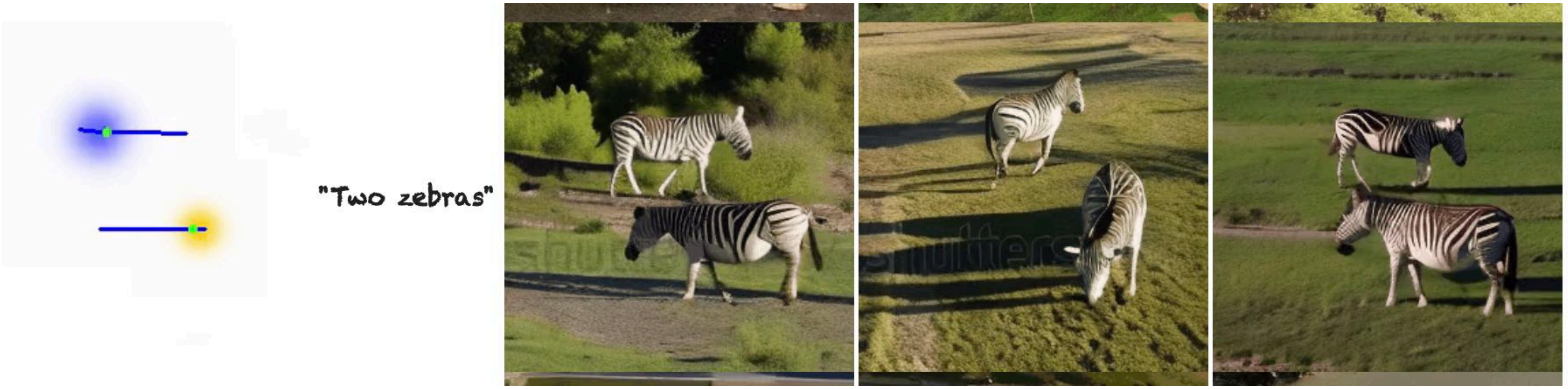
Training data annotation (for video)

What type of annotations might we want?

For text conditioning: text description of each video:

But are properties of a “good” description? *

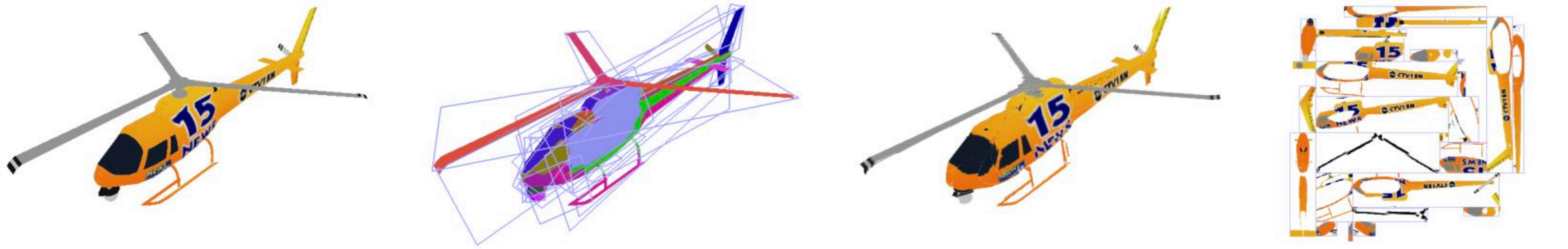
For movement vector conditioning: key object detections and vectors of movement of these objects



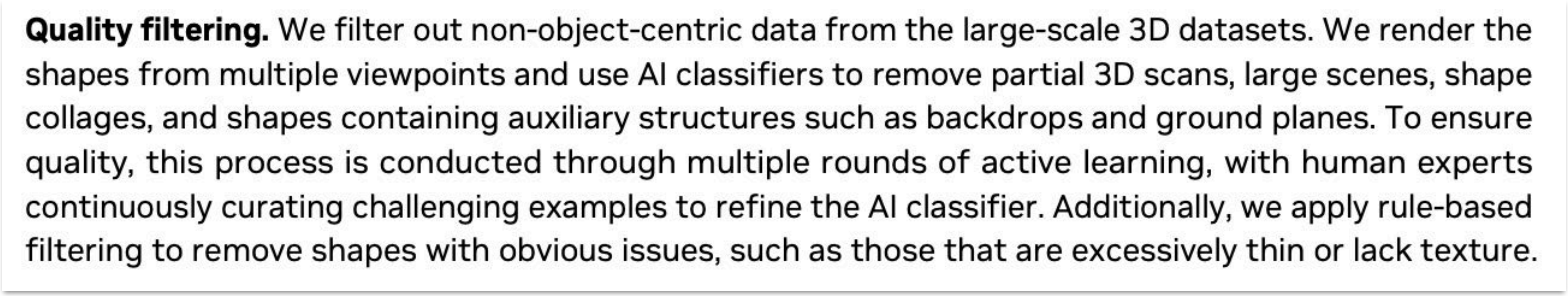
* We’ll dive deeper into this later in the slides.

What about curating 3D meshes for 3D generation?

- Similar curation issues
  - Detect inappropriate/duplicate/copyrighted
    - Render meshes and reduce to curation of images?
    - Directly analyze geometry?
  - Detect meshes with undesirable 3D properties
    - Simple geometry: simple cubes and planes used as “billboards” (that are intended to have texture)
    - “Low-quality” models with holes
    - Fine/thin structures that current models can’t reproduce
    - Many widely used “big” 3D datasets contain vast amounts of data that are filtered prior to training

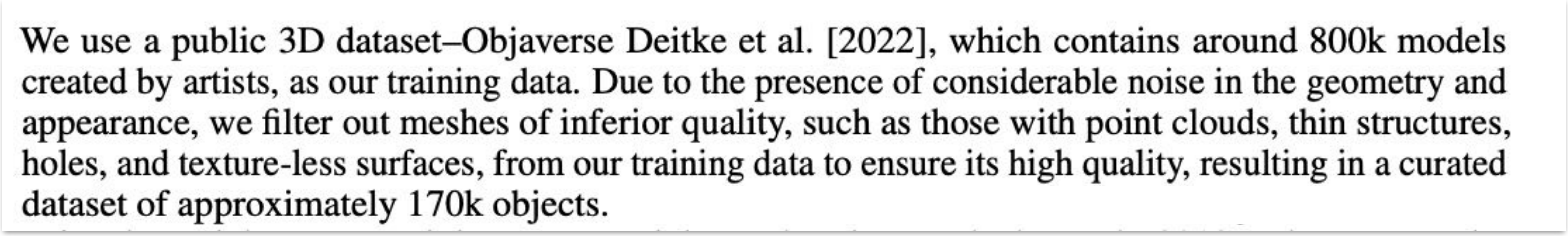


[Image credit: Décoret et al. 2023]



Quality filtering. We filter out non-object-centric data from the large-scale 3D datasets. We render the shapes from multiple viewpoints and use AI classifiers to remove partial 3D scans, large scenes, shape collages, and shapes containing auxiliary structures such as backdrops and ground planes. To ensure quality, this process is conducted through multiple rounds of active learning, with human experts continuously curating challenging examples to refine the AI classifier. Additionally, we apply rule-based filtering to remove shapes with obvious issues, such as those that are excessively thin or lack texture.

From Edify 3D, [NVIDIA 2024]

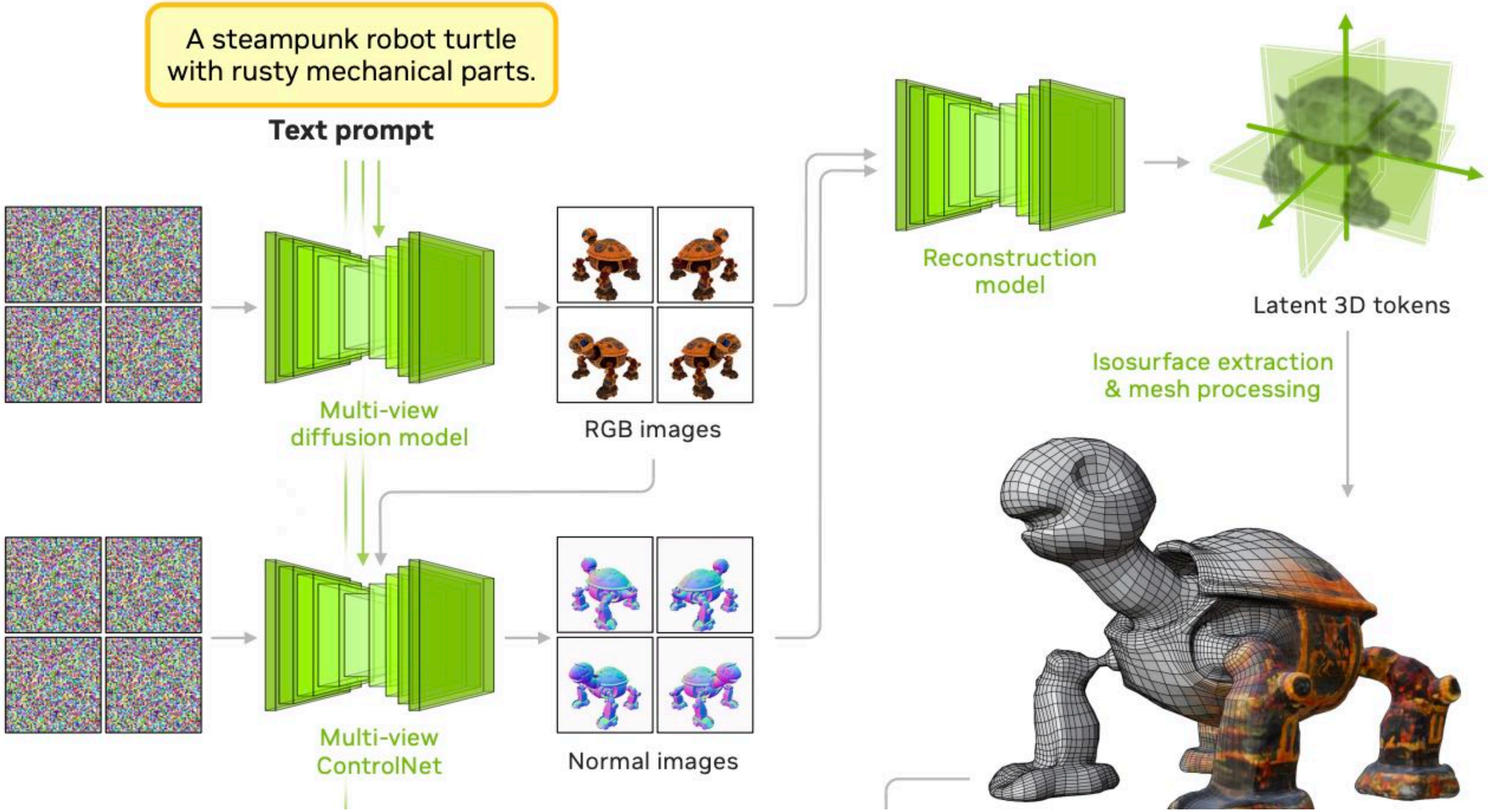


We use a public 3D dataset–Objaverse Deitke et al. [2022], which contains around 800k models created by artists, as our training data. Due to the presence of considerable noise in the geometry and appearance, we filter out meshes of inferior quality, such as those with point clouds, thin structures, holes, and texture-less surfaces, from our training data to ensure its high quality, resulting in a curated dataset of approximately 170k objects.

From Craftsman, [Li et al. 2024]

Generating source content for 3D generation

- **Lack of good 3D models has led to interest in leveraging good 2D image generation models to generate 3D content for use as training data.**


A steampunk robot turtle with rusty mechanical parts.
Text prompt
Multi-view diffusion model
RGB images
Multi-view ControlNet
Normal images
Reconstruction model
Latent 3D tokens
Isosurface extraction & mesh processing

From Edify 3D, [NVIDIA 2024]

How to train a model to produce different views of the same object?

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

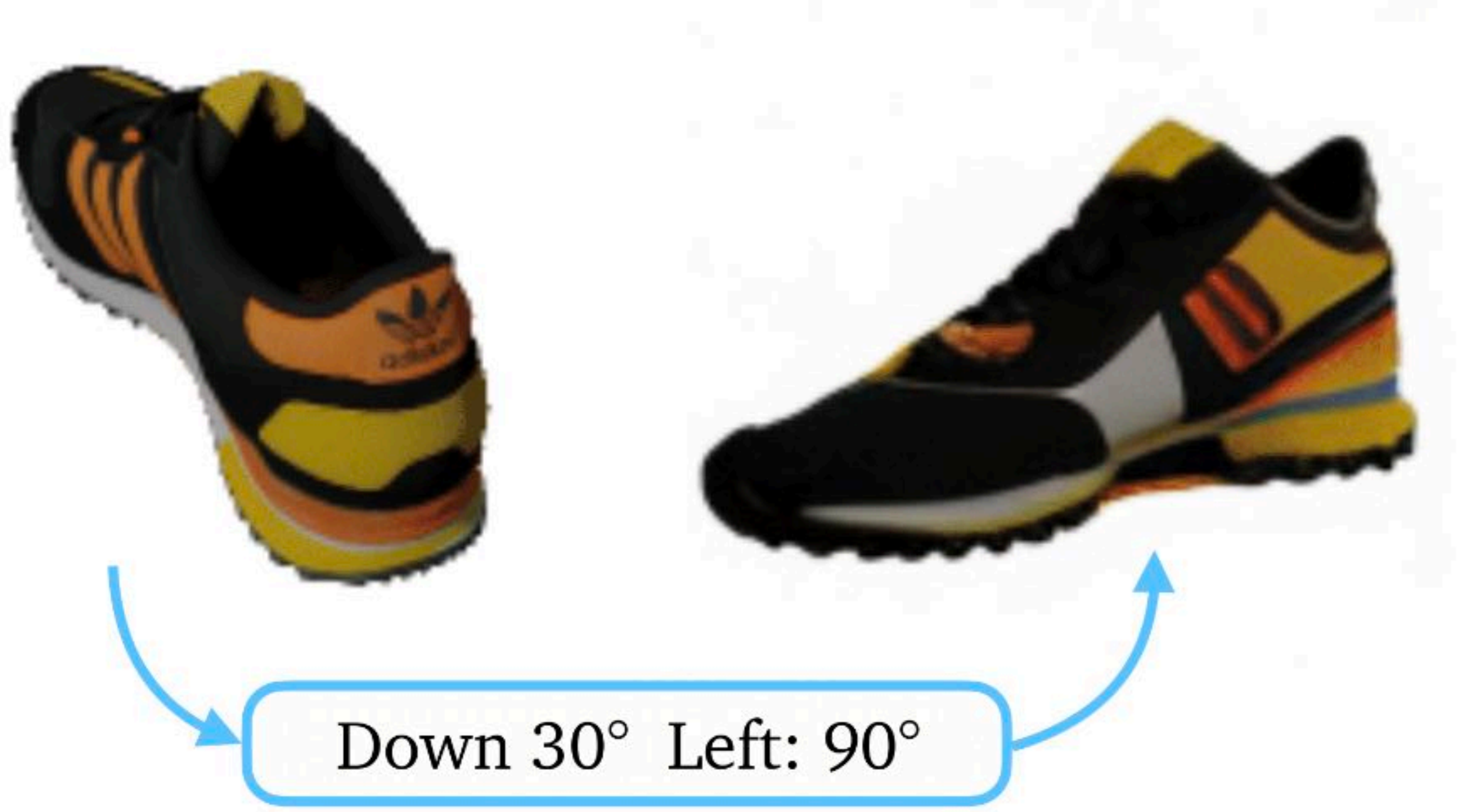
Rendered from viewpoint 1

Rendered from viewpoint 2

Input (conditioning)

Output

(Object, camera info)



Automatically synthesizing text descriptions that lead to models that exhibit good alignment between text input and generated output assets is a hard problem

(Note: this is about improving text-based controls)

Common ideas in caption generation

- Leverage modern VLMs to generate captions from images
  - *Huge amounts of proprietary prompt engineering*
  - In case of 3D model generation, render model to get the image (from what views?)
- Generate captions at multiple levels of detail
  - Provides many different descriptions of the same asset to the model (reduce overfitting)
  - Captures what user might prompt at different levels of precision
    - For short captions, creates one-to-many relationship between short captions and assets so that given a short-general-text description, model learns diverse distribution of answers
    - For detailed captions, model learns more precise mapping of parts of caption to details of asset

Caption_1: A close-up photograph of a dish featuring a stuffed acorn squash filled with couscous and cheese, garnished with fresh cranberries and chopped parsley. The squash is placed on a white plate, and the dish is set on a green table with additional cranberries and parsley scattered around. In the background, there is a block of cheese with the label 'PARSLEY CHIVES' partially visible. The lighting is soft and even, highlighting the vibrant colors of the dish and the fresh ingredients. The composition is focused on the squash, with the background elements serving to enhance the overall presentation.

Caption_2: A close-up of a dish featuring a stuffed acorn squash with couscous and cheese, garnished with cranberries and parsley. The squash is on a white plate on a green table with additional cranberries and parsley scattered around.

Caption_3: A stuffed acorn squash on a white plate, garnished with cranberries and parsley, on a green table with additional ingredients scattered around.

Caption_4: Stuffed acorn squash on a plate, garnished with cranberries and parsley, on a green table.

Caption_5: Stuffed acorn squash on a plate.

Caption_6: Squash dish.

Caption_1: A serene and picturesque scene captured through a window looking out onto a lush garden and majestic mountains. The foreground is dominated by a variety of vibrant flowers in pink, purple, and white hues, arranged in a cascading manner. The garden is well-maintained with green grass, a few trees, and a wooden fence. In the midground, a well-maintained lawn extends towards the background, bordered by a few tall, slender trees. The background features a breathtaking view of a mountain range with snow-capped peaks, partially obscured by clouds. The sky above is clear with a few scattered clouds, allowing for natural daylight to illuminate the entire scene. The overall composition is balanced, with the vibrant flowers in the foreground adding a pop of color to the serene mountainous backdrop.

Caption_2: A serene scene captured through a window featuring a lush garden and majestic mountains. Vibrant flowers in pink and purple hues dominate the foreground, with a well-maintained lawn and trees in the midground. The background shows a breathtaking mountain range with snow-capped peaks.

Caption_3: A peaceful scene through a window, with colorful flowers in the foreground, a well-maintained lawn, and majestic mountains in the background.

Caption_4: A window view of a garden with vibrant flowers and a mountainous backdrop.

Caption_5: Garden and mountains through a window.

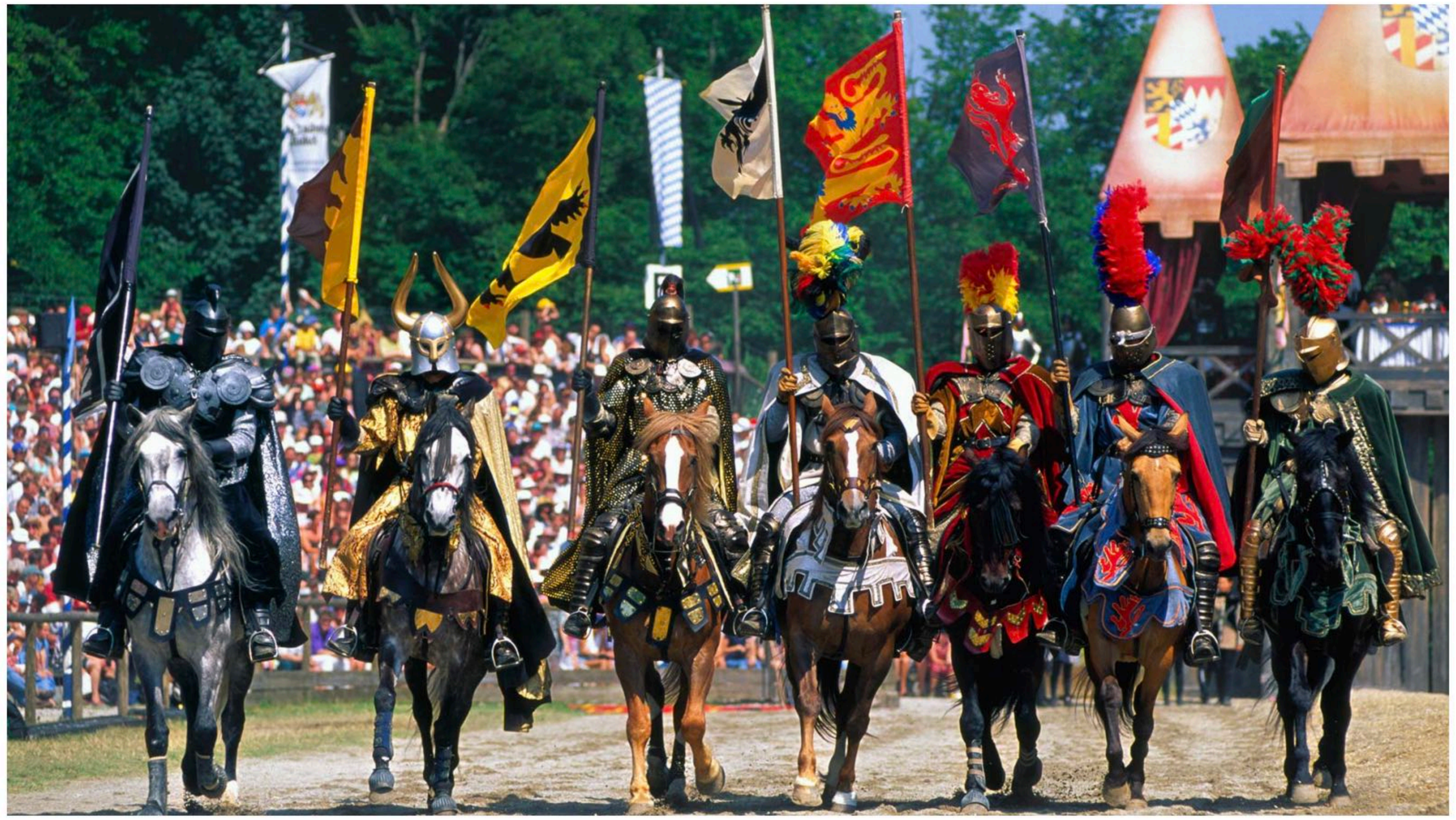
Caption_6: Nature scene.

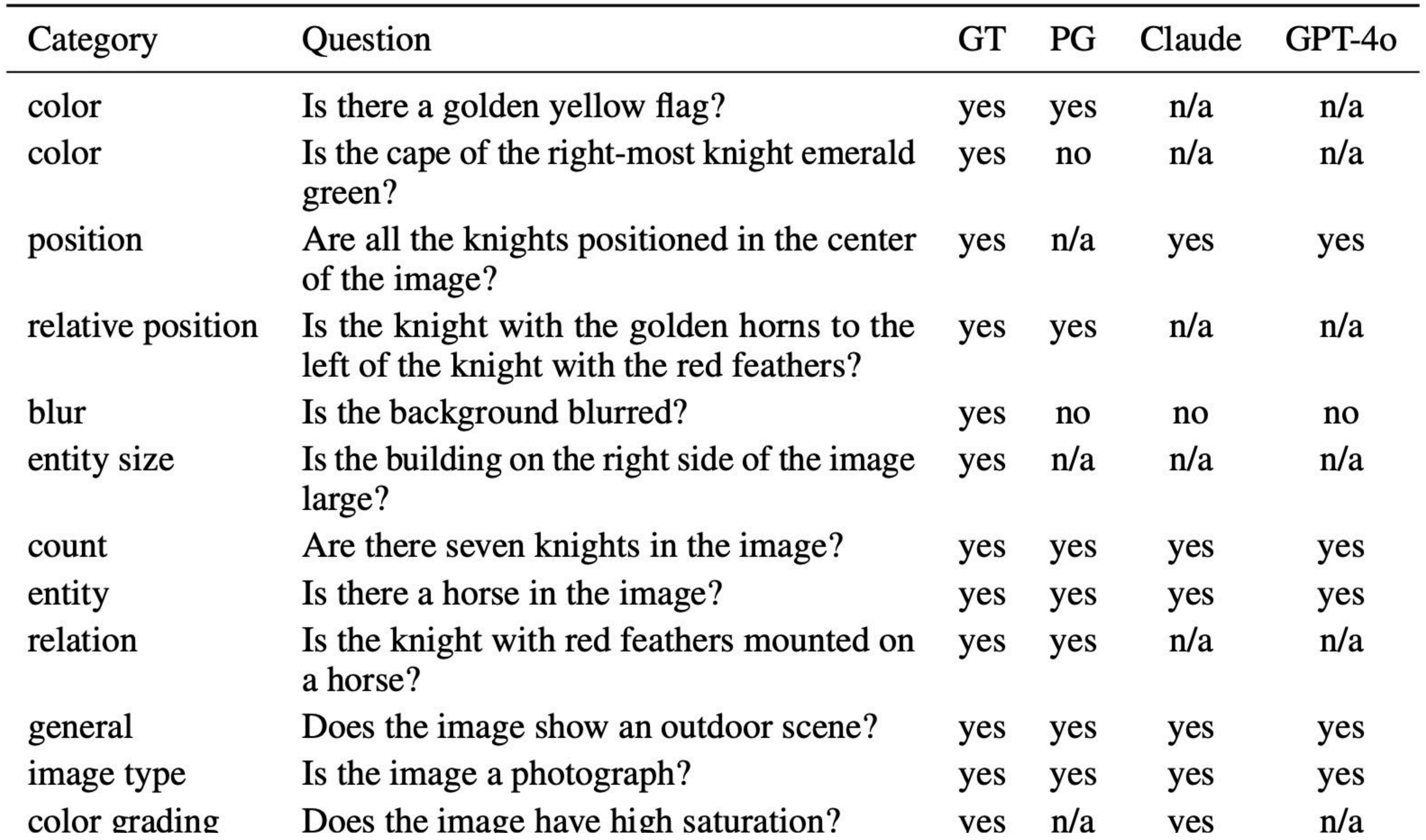
Source: Playground V3 [Lui et al. 2024]

How do we validate the “quality” of these automatically generated captions?

Automated caption evaluation strategies

- General approach: check consistency of text captions with results of direct image analysis
- Run object detector on image, see if objects detected are mentioned in the caption
- Is embedding (CLIP score) of image similar to embedding of caption?
- Can an LLM (given only the caption) come up with the same answers to questions about an image as a VLM (given the image)?
  - How do we generate image-specific questions?
- For 3D models: are captions produced from rendered images consistent across different renderings of the same model?
  - What should the views be?

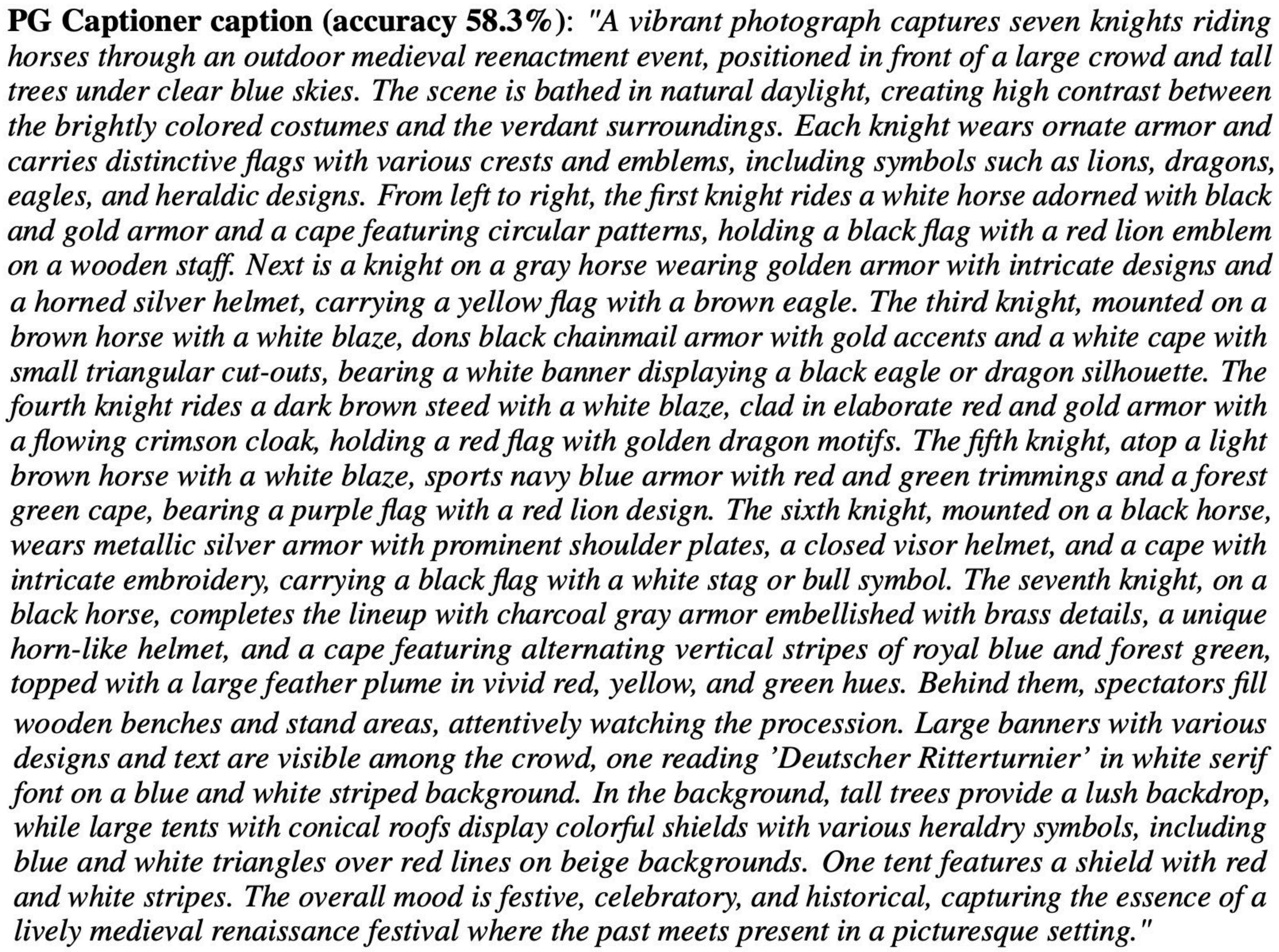




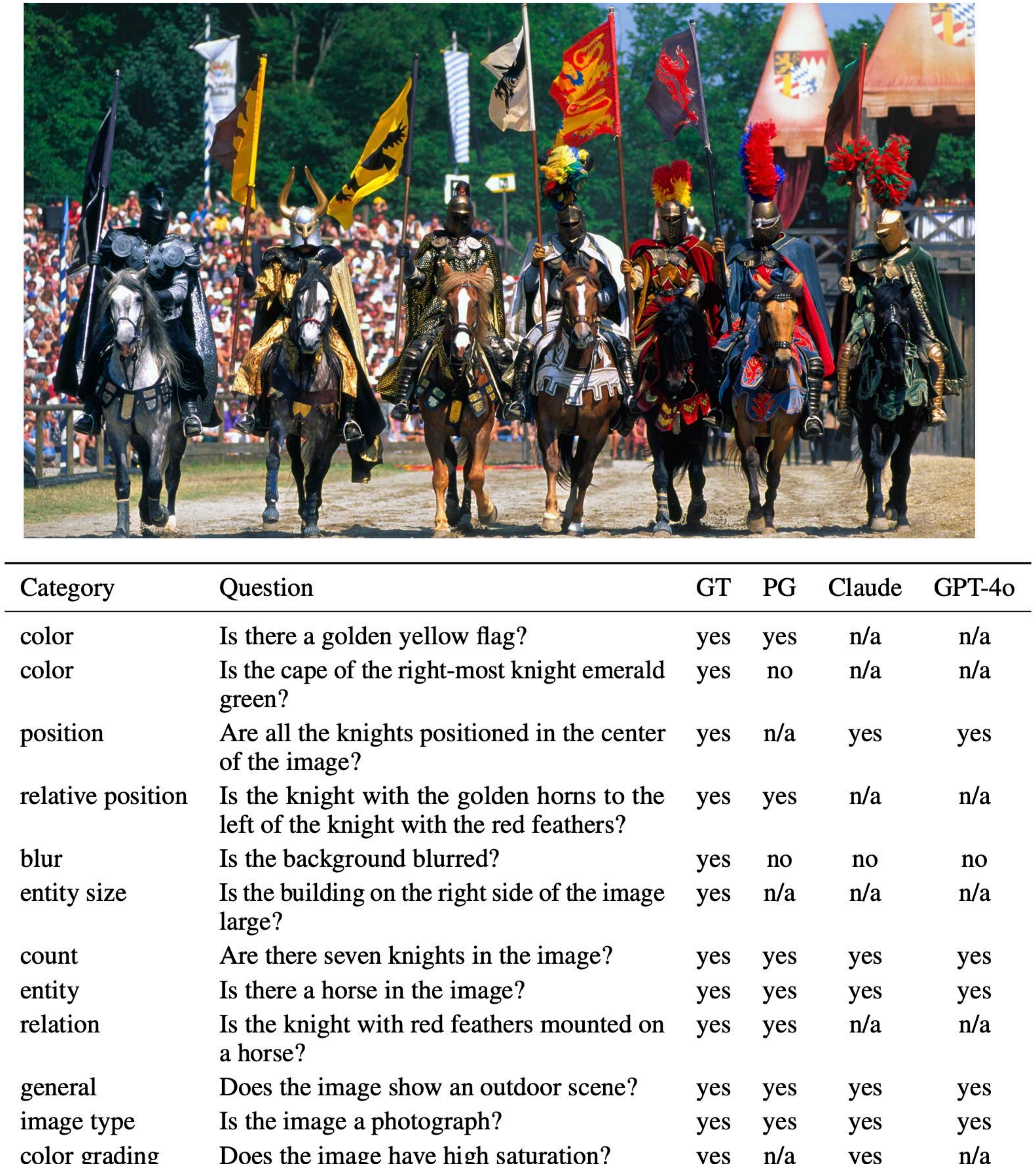
Category	Question	GT	PG	Claude	GPT-4o
color	Is there a golden yellow flag?	yes	yes	n/a	n/a
color	Is the cape of the right-most knight emerald green?	yes	no	n/a	n/a
position	Are all the knights positioned in the center of the image?	yes	n/a	yes	yes
relative position	Is the knight with the golden horns to the left of the knight with the red feathers?	yes	yes	n/a	n/a
blur	Is the background blurred?	yes	no	no	no
entity size	Is the building on the right side of the image large?	yes	n/a	n/a	n/a
count	Are there seven knights in the image?	yes	yes	yes	yes
entity	Is there a horse in the image?	yes	yes	yes	yes
relation	Is the knight with red feathers mounted on a horse?	yes	yes	n/a	n/a
general	Does the image show an outdoor scene?	yes	yes	yes	yes
image type	Is the image a photograph?	yes	yes	yes	yes
color grading	Does the image have high saturation?	yes	n/a	yes	n/a

Source: Playground V3 [Lui et al. 2024]

Caption evaluation ideas



PG Captioner caption (accuracy 58.3%): *"A vibrant photograph captures seven knights riding horses through an outdoor medieval reenactment event, positioned in front of a large crowd and tall trees under clear blue skies. The scene is bathed in natural daylight, creating high contrast between the brightly colored costumes and the verdant surroundings. Each knight wears ornate armor and carries distinctive flags with various crests and emblems, including symbols such as lions, dragons, eagles, and heraldic designs. From left to right, the first knight rides a white horse adorned with black and gold armor and a cape featuring circular patterns, holding a black flag with a red lion emblem on a wooden staff. Next is a knight on a gray horse wearing golden armor with intricate designs and a horned silver helmet, carrying a yellow flag with a brown eagle. The third knight, mounted on a brown horse with a white blaze, dons black chainmail armor with gold accents and a white cape with small triangular cut-outs, bearing a white banner displaying a black eagle or dragon silhouette. The fourth knight rides a dark brown steed with a white blaze, clad in elaborate red and gold armor with a flowing crimson cloak, holding a red flag with golden dragon motifs. The fifth knight, atop a light brown horse with a white blaze, sports navy blue armor with red and green trimmings and a forest green cape, bearing a purple flag with a red lion design. The sixth knight, mounted on a black horse, wears metallic silver armor with prominent shoulder plates, a closed visor helmet, and a cape with intricate embroidery, carrying a black flag with a white stag or bull symbol. The seventh knight, on a black horse, completes the lineup with charcoal gray armor embellished with brass details, a unique horn-like helmet, and a cape featuring alternating vertical stripes of royal blue and forest green, topped with a large feather plume in vivid red, yellow, and green hues. Behind them, spectators fill wooden benches and stand areas, attentively watching the procession. Large banners with various designs and text are visible among the crowd, one reading 'Deutscher Ritterturnier' in white serif font on a blue and white striped background. In the background, tall trees provide a lush backdrop, while large tents with conical roofs display colorful shields with various heraldry symbols, including blue and white triangles over red lines on beige backgrounds. One tent features a shield with red and white stripes. The overall mood is festive, celebratory, and historical, capturing the essence of a lively medieval renaissance festival where the past meets present in a picturesque setting."*



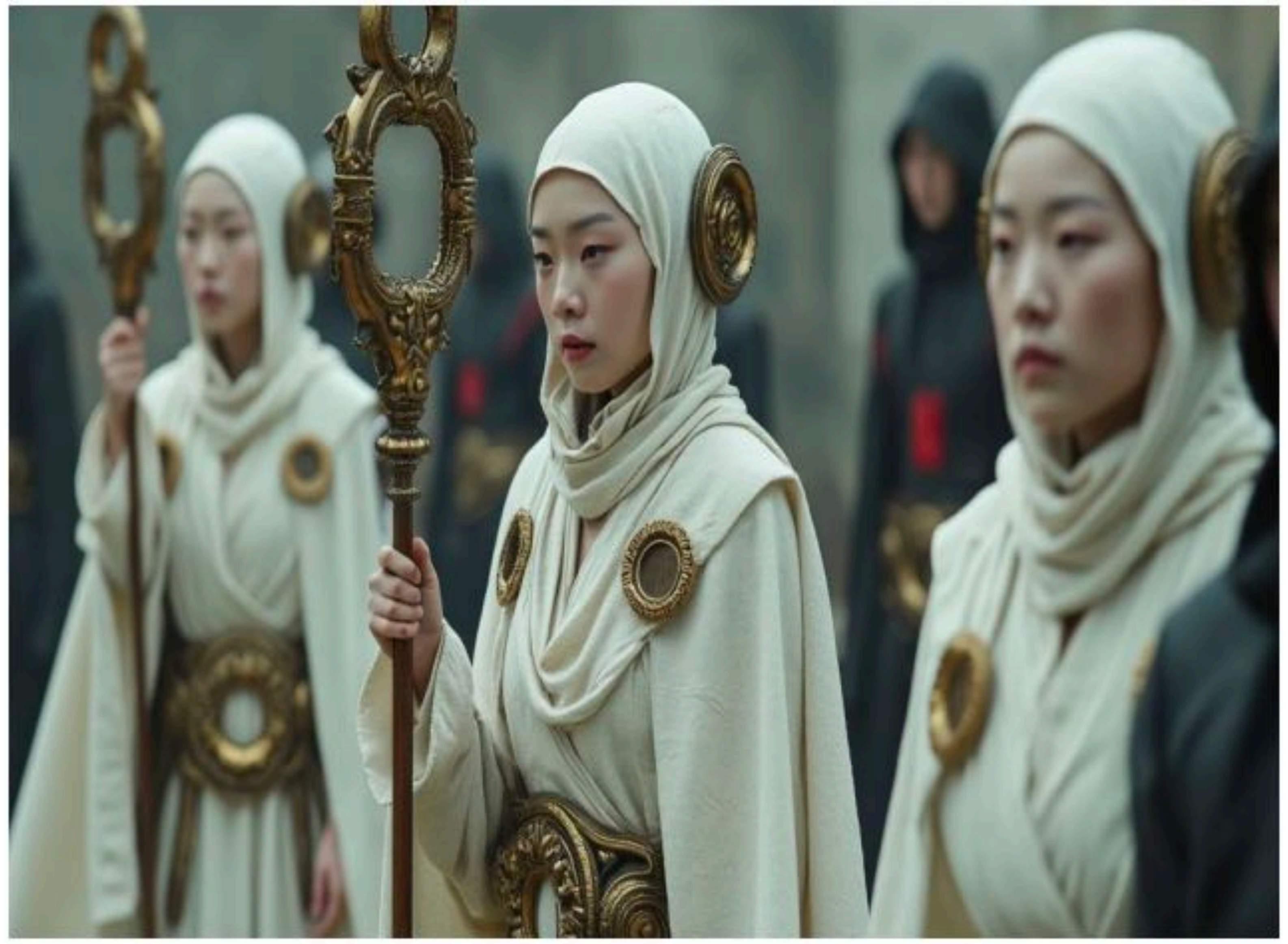
Category	Question	GT	PG	Claude	GPT-4o
color	Is there a golden yellow flag?	yes	yes	n/a	n/a
color	Is the cape of the right-most knight emerald green?	yes	no	n/a	n/a
position	Are all the knights positioned in the center of the image?	yes	n/a	yes	yes
relative position	Is the knight with the golden horns to the left of the knight with the red feathers?	yes	yes	n/a	n/a
blur	Is the background blurred?	yes	no	no	no
entity size	Is the building on the right side of the image large?	yes	n/a	n/a	n/a
count	Are there seven knights in the image?	yes	yes	yes	yes
entity	Is there a horse in the image?	yes	yes	yes	yes
relation	Is the knight with red feathers mounted on a horse?	yes	yes	n/a	n/a
general	Does the image show an outdoor scene?	yes	yes	yes	yes
image type	Is the image a photograph?	yes	yes	yes	yes
color grading	Does the image have high saturation?	yes	n/a	yes	n/a

Source: Playground V3 [Lui et al. 2024]

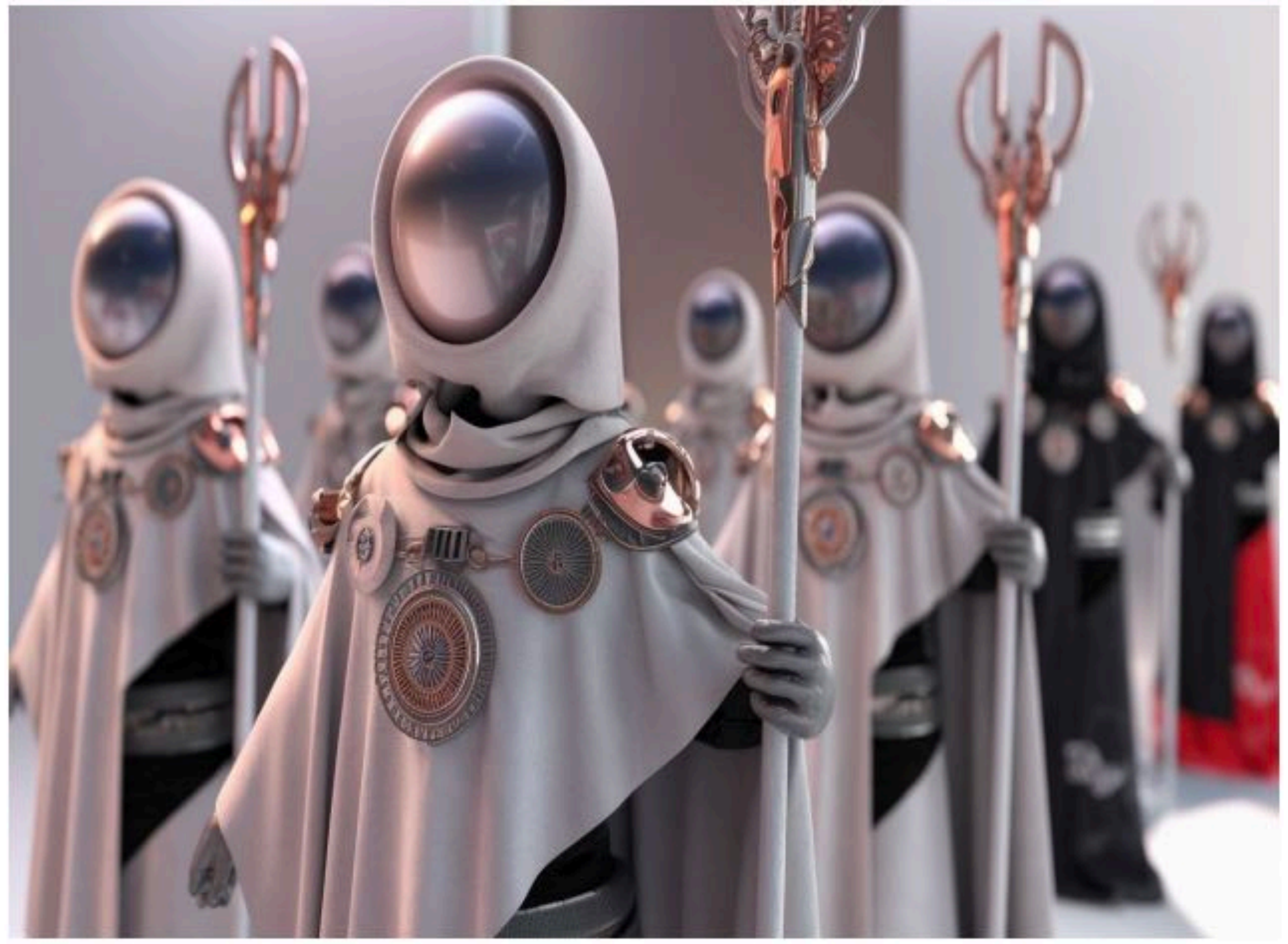
Impact of captions

- Output of three different models trained on different captions
  - Bold colored text highlights places where models 1 and 2 fail to adhere to prompt

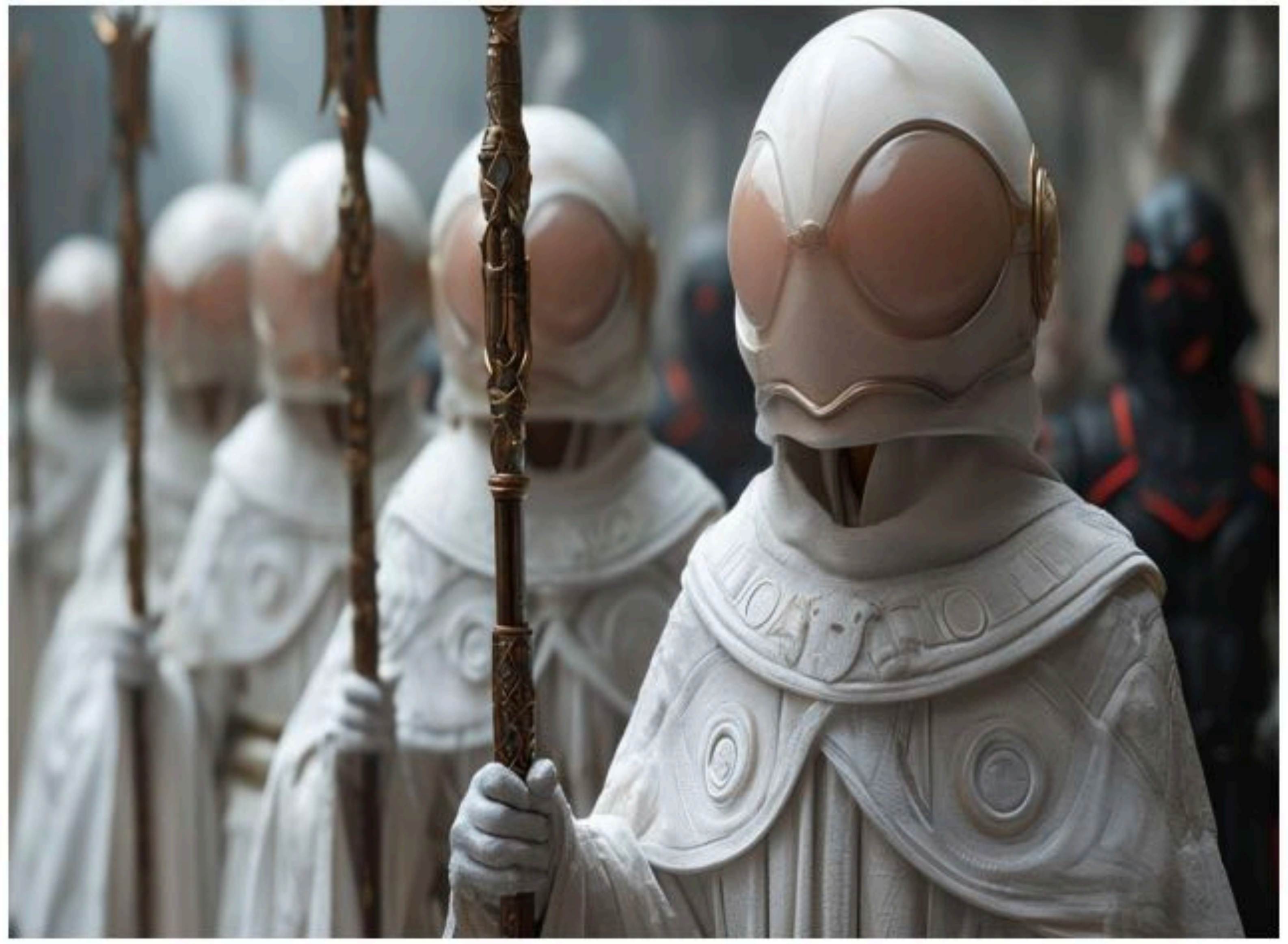
Model 1



Model 2



Model 3



Prompt:

A cinematic scene featuring a group of figures dressed in ornate, white robes with intricate circular patterns and futuristic elements.
The figures have large, dome-shaped helmets with a soft, pinkish hue, obscuring their faces.
They are holding staffs with detailed bronze and gold designs, adding a ceremonial or ritualistic feel to the scene.
The foreground figure, positioned slightly to the right, is in sharp focus, showcasing the texture and design of the robe and staff.
The midground features additional figures, also in focus but slightly blurred, creating depth.
In the background, there are indistinct figures in dark clothing with red accents, adding a sense of mystery and depth.
The lighting is soft and diffused, creating a balanced and mysterious atmosphere.
The color palette includes white, light gray, bronze, gold, black, and red, contributing to the scene's otherworldly and ceremonial mood.

Takeaways

- **Quality of techniques for automatically selecting and annotating training data for generative AI can have great impact on the usefulness of end models ("usefulness" = depends on both quality of generations AND ability to control generation in useful ways)**

- **Significant use of both conventional analysis tools and AI-based tools to perform data filtering and annotation**

- **The difference in model performance between "good" and "bad" datasets can be substantial**
  - **As big or bigger than differences between different model architectures, or models of different size.**
  - ***Notice that companies are willing to publish source code for their models and write papers on the DNN architectures, but they publish far fewer details on the specifics of their data curation pipelines.*** 🤔🤔🤔

- **Difficult to evaluate dataset quality, and to attribute dataset selection choices to changes in end model performance**
  - **It's currently a black art… could use more systematic research to understand better**
  - **Tonight's reading: early systematic studies for dataset selection for text generation models (LLMs)**