

Working with Unpredictable Black Boxes

CS 347

Maneesh Agrawala

Announcements

Quiz 3 a week from today

Last time

Intelligence augmentation aims to place AI in context by using it to amplify our own abilities

Debates rage about the levels of autonomy to grant to AIs: from fully autonomous **agents** that act on the person's behalf, to **direct manipulation** that always leaves the user in full control

Mixed initiative interaction splits the difference by asking, acting, or doing nothing based on its confidence and assessment of the benefit

End users and designers seek to work with these AI tools

Today

Problem:

Unpredictable tools are terrible interfaces

Why? Because we have no good conceptual model

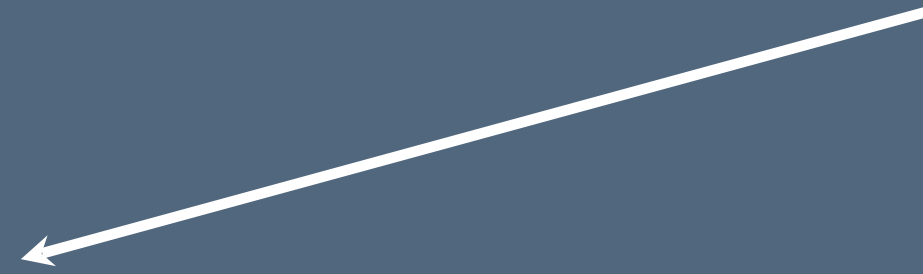
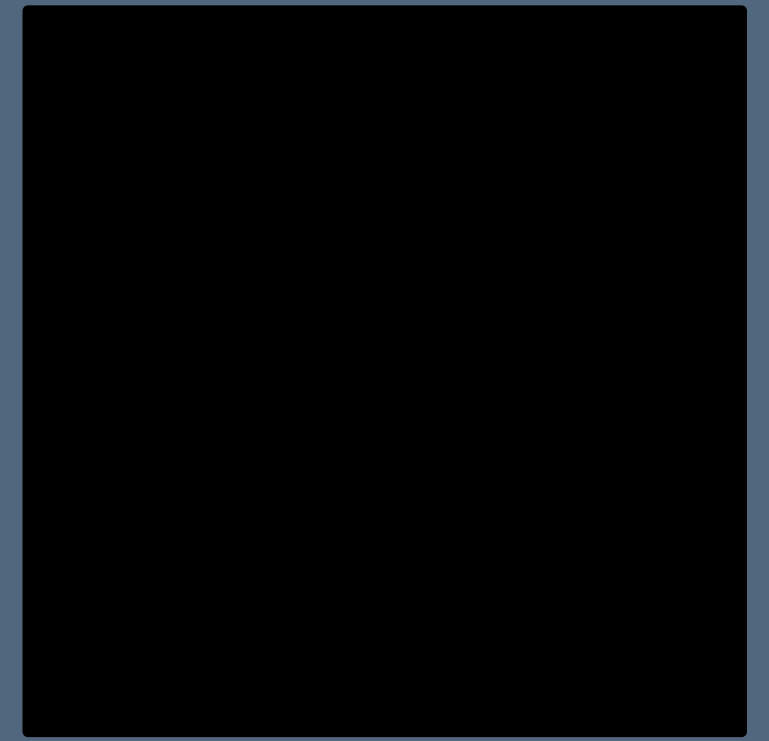
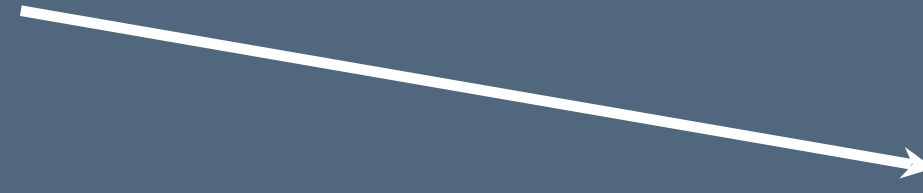
Solutions:

Towards conversational AI interfaces

Dealing with ambiguity of natural language

Iterative refinement (not iterative trial-and-error)

Picture of a Professor named Maneesh Agrawala



DALL-E2





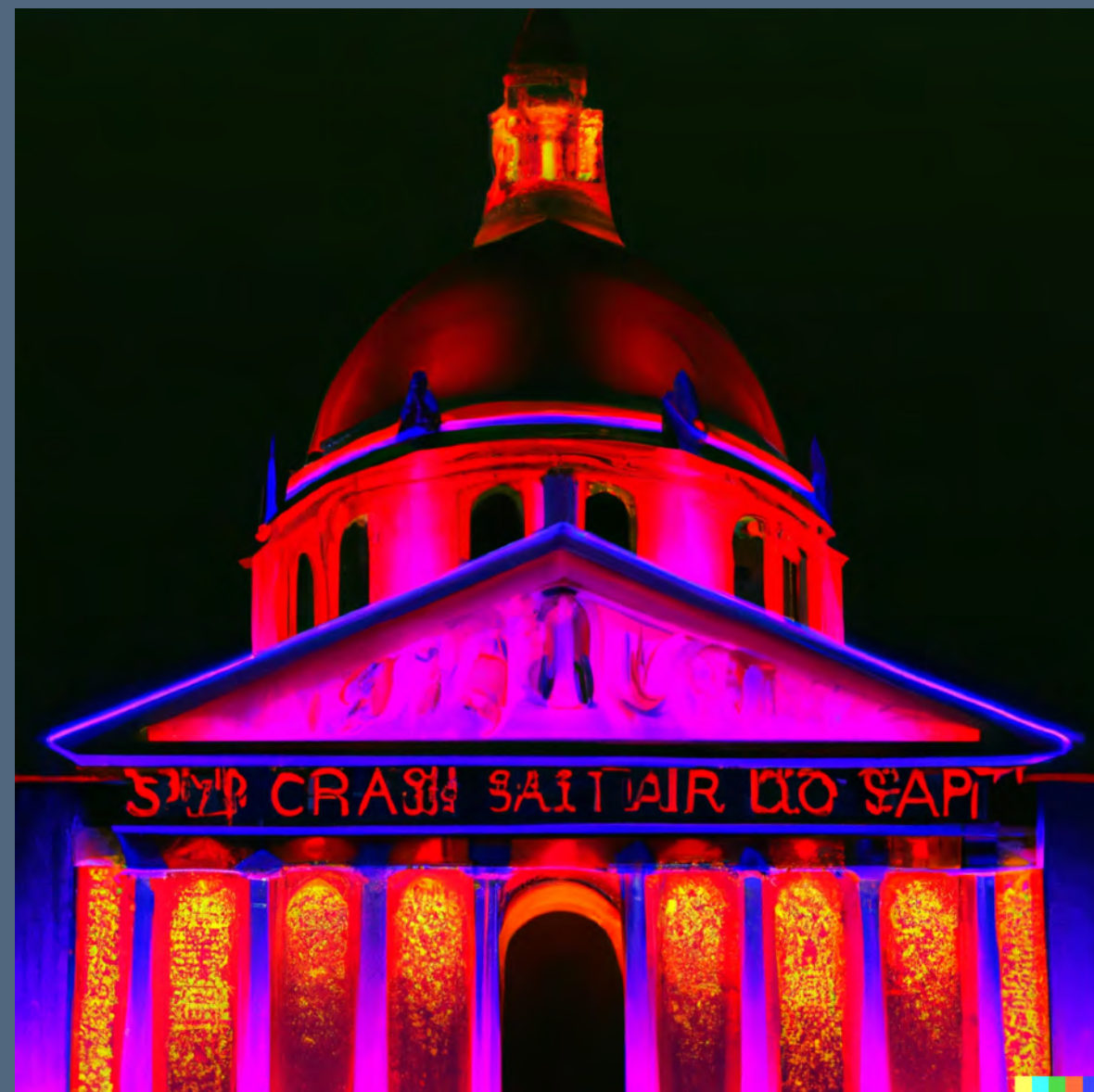


stanford memorial church with neon signage in the style of bladerunner

stanford memorial church **and main quad with palm trees** in the style of bladerunner

nighttime rain stanford memorial church and main quad with palm trees, **night market food stalls and neon signs** in the style of bladerunner

nighttime rain stanford memorial church and main quad with palm trees, night market food stalls and neon signs **like downtown tokyo**



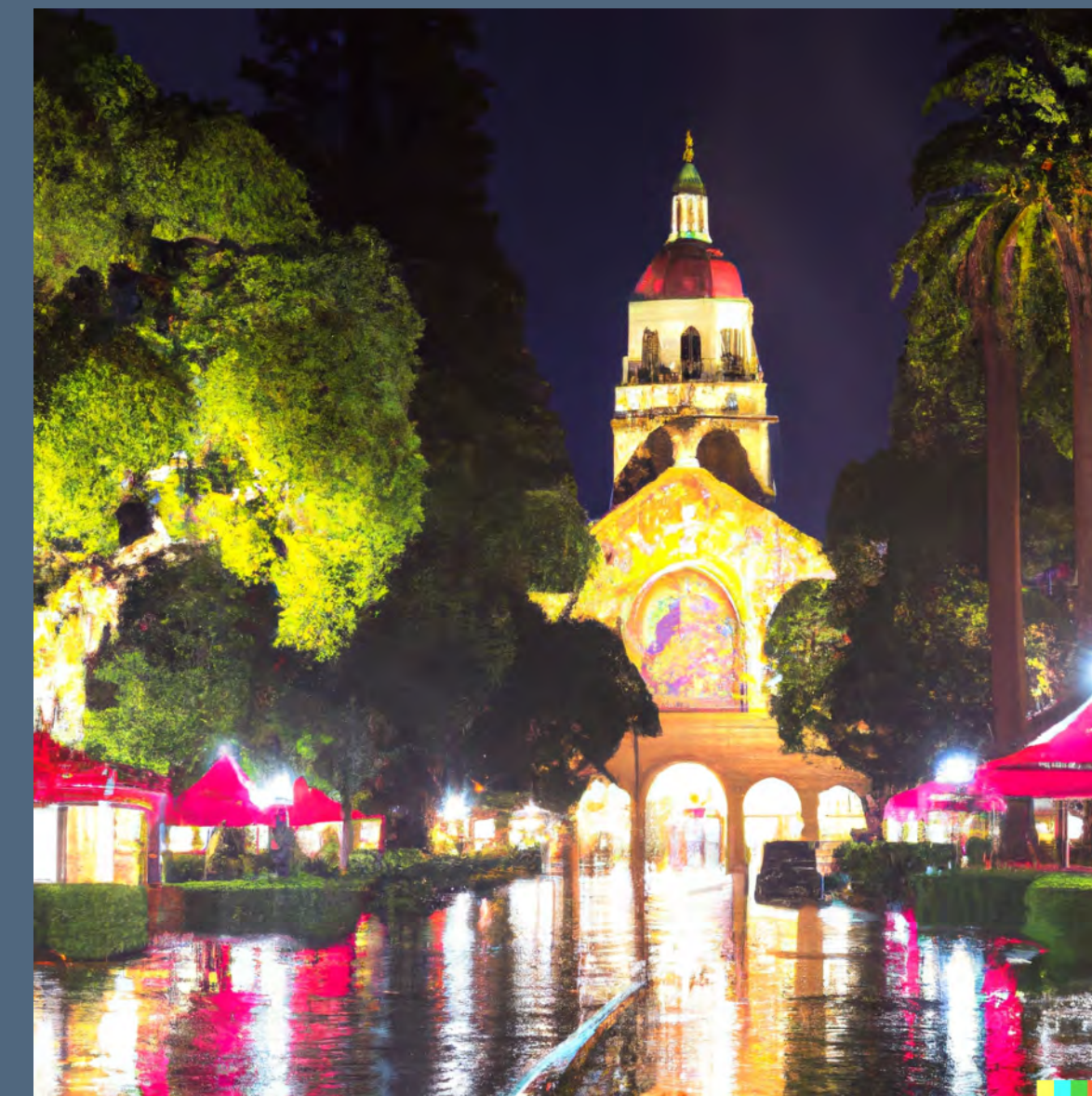
Iteration 1



Iteration 3



Iteration 8



Iteration 17

← Back

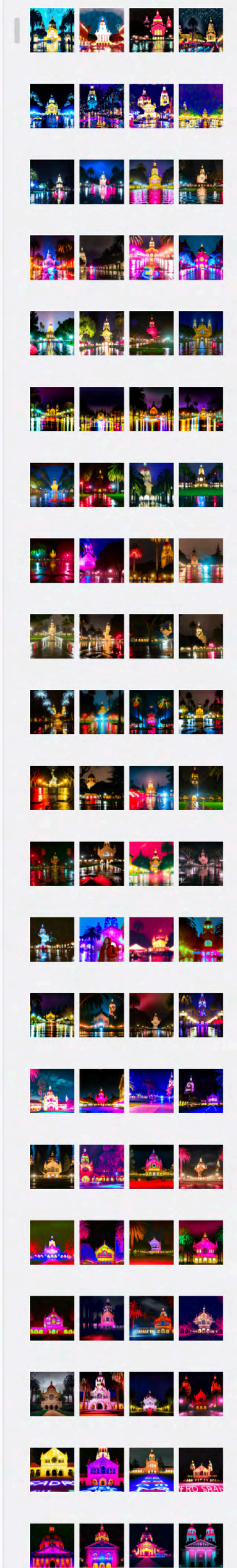
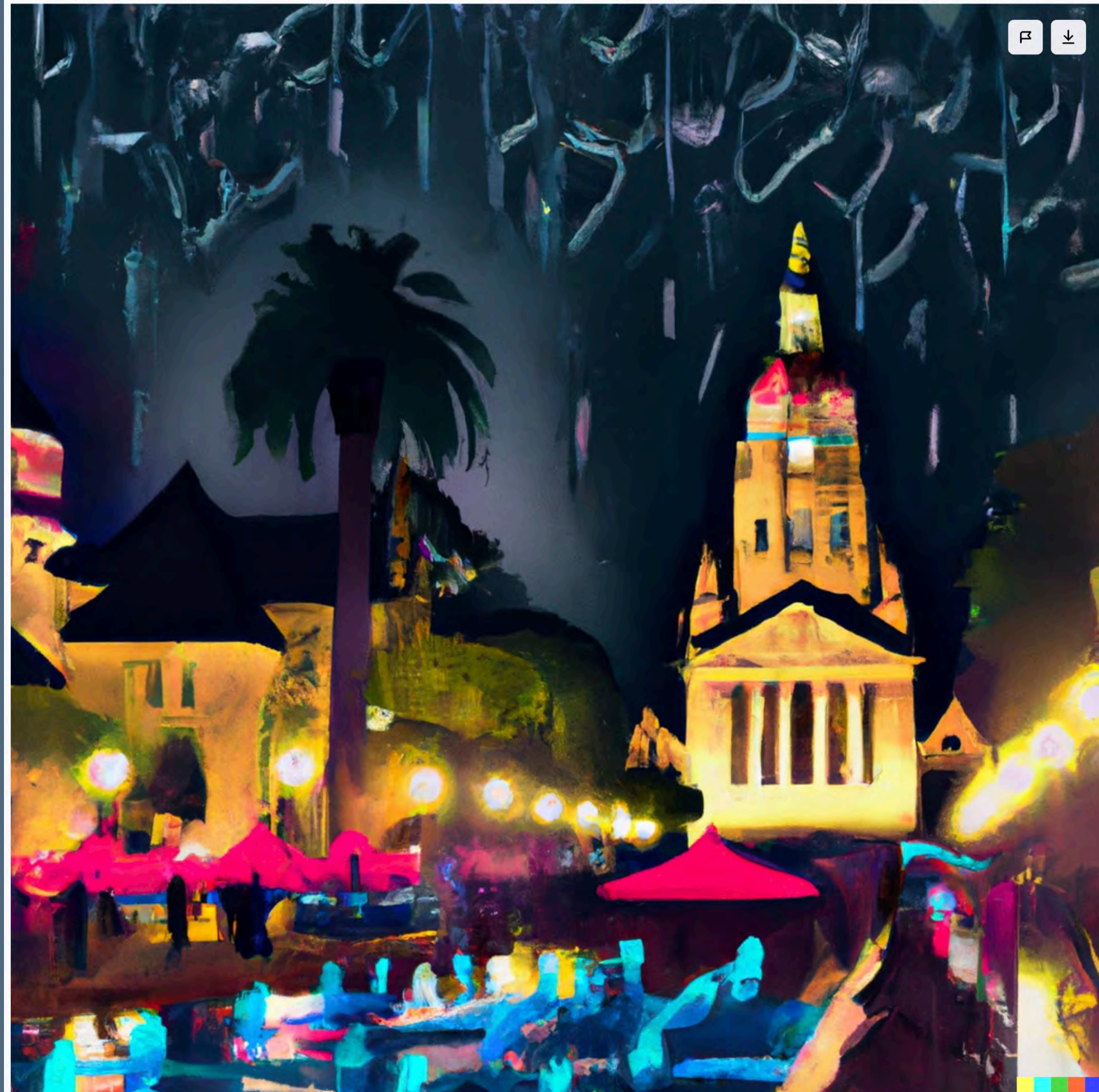
Edit

Variations

Share

Save ▾

→



nighttime rain stanford
memorial church and
main quad with palm
trees, night market
japadog food stalls and
neon signs, **neo** tokyo
bladerunner style **film**
still illustration

Iteration 21



Aaron Hertzmann

@AaronHertzmann



Writing a letter and quite happy with this phrase: Real artistic tools should act as extensions of the artist, the way a paintbrush adds capabilities to a painter's hand, rather than a slot machine that may or may not give you something useful.

8:05 AM · Sep 25, 2023 · **5,562** Views



Beautiful Watercolor Illustrations

505  @imagineer

Generates beautiful watercolor illustrations with undefined figures, in a consistent style. Ideal for illustrating stories, tales or blogs with vivid and colorful watercolor images. You can select the proportions of each generated illustration.

\$1.99

Get Prompt

After purchasing, you will gain access to the prompt file, which you can use with Midjourney. You must already have access to Midjourney to use this prompt.



Why Johnny Can't Prompt

[Zamfirescu-Pereira et al. 2023]

YOU READ THIS

Prompters **don't know what AI can/cannot do**. So need examples or instructions on how to proceed. Consistent with [Yang 2020].

Prompters **over-generalize** from a few examples, or errors (give up early).

Prompters **anthropomorphize** and filter expectations based on human-human interactions.

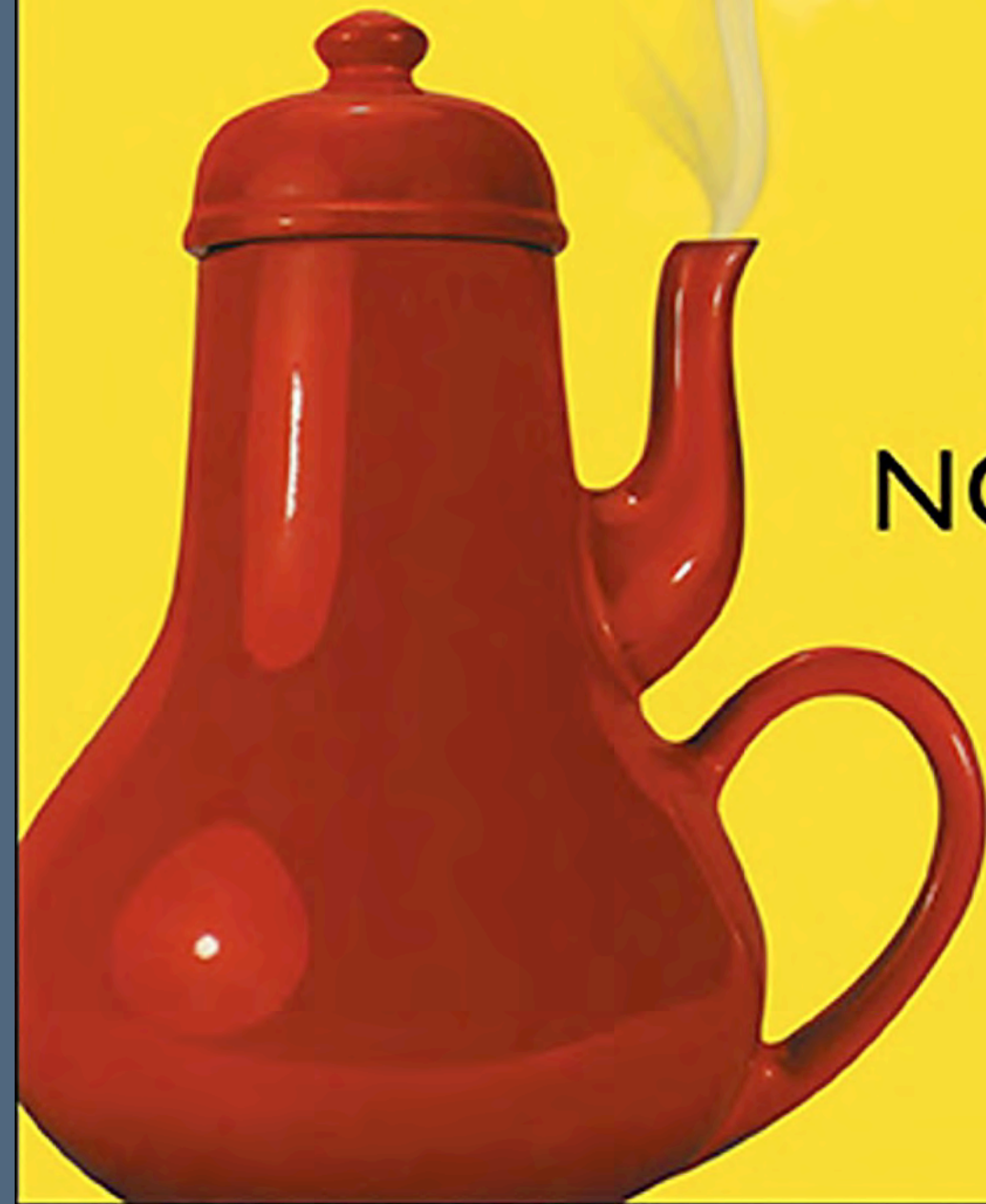
Gave direct instructions instead of providing in-context examples. Even when instructed by human researcher to give examples.

Some prompters expected AI to understand instructions the way a human would (e.g. instruction: 'do not use ABC', result: AI uses ABC verbatim in response)

Conceptual Models

REVISED & EXPANDED EDITION

The DESIGN
of EVERYDAY
THINGS



DON
NORMAN



freezer



fresh food



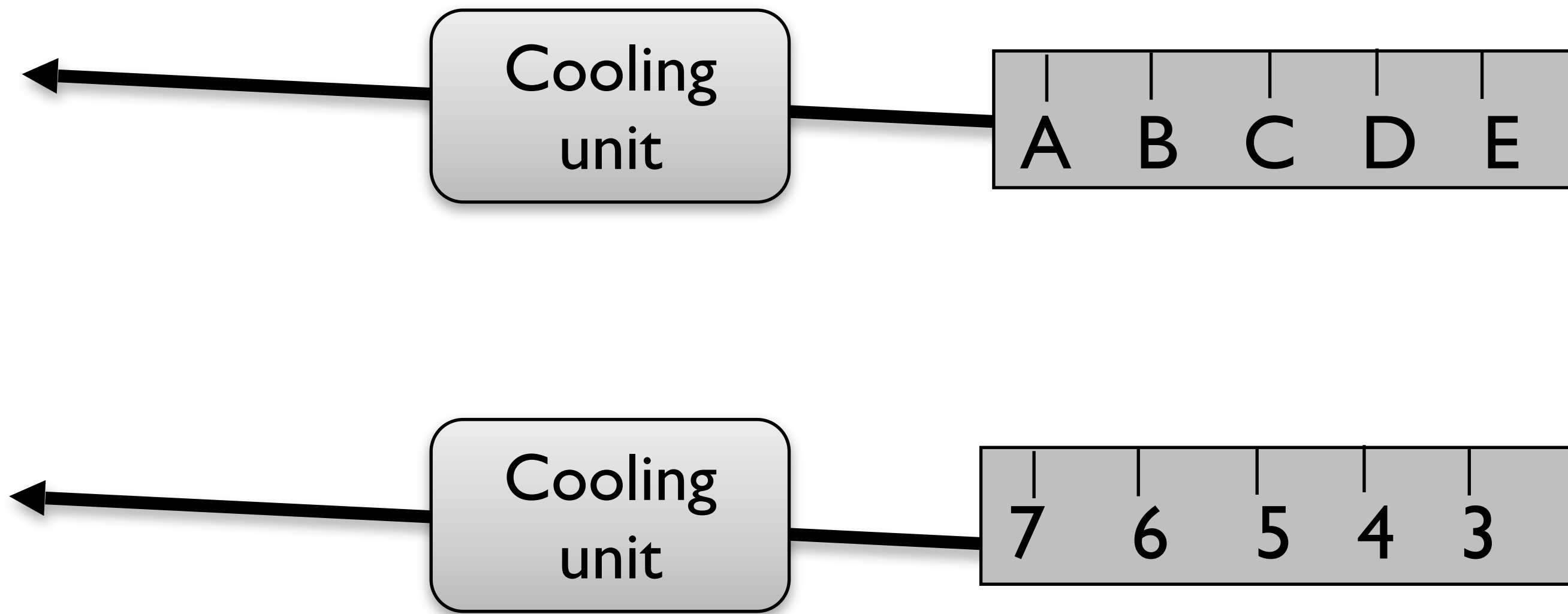
Normal Settings
Colder Fresh Food
Coldest Fresh Food
Colder Freezer
Warmer Fresh Food
OFF (both)

C and 4
C and 5-6
B and 7
D and 6-7
C and 3-1

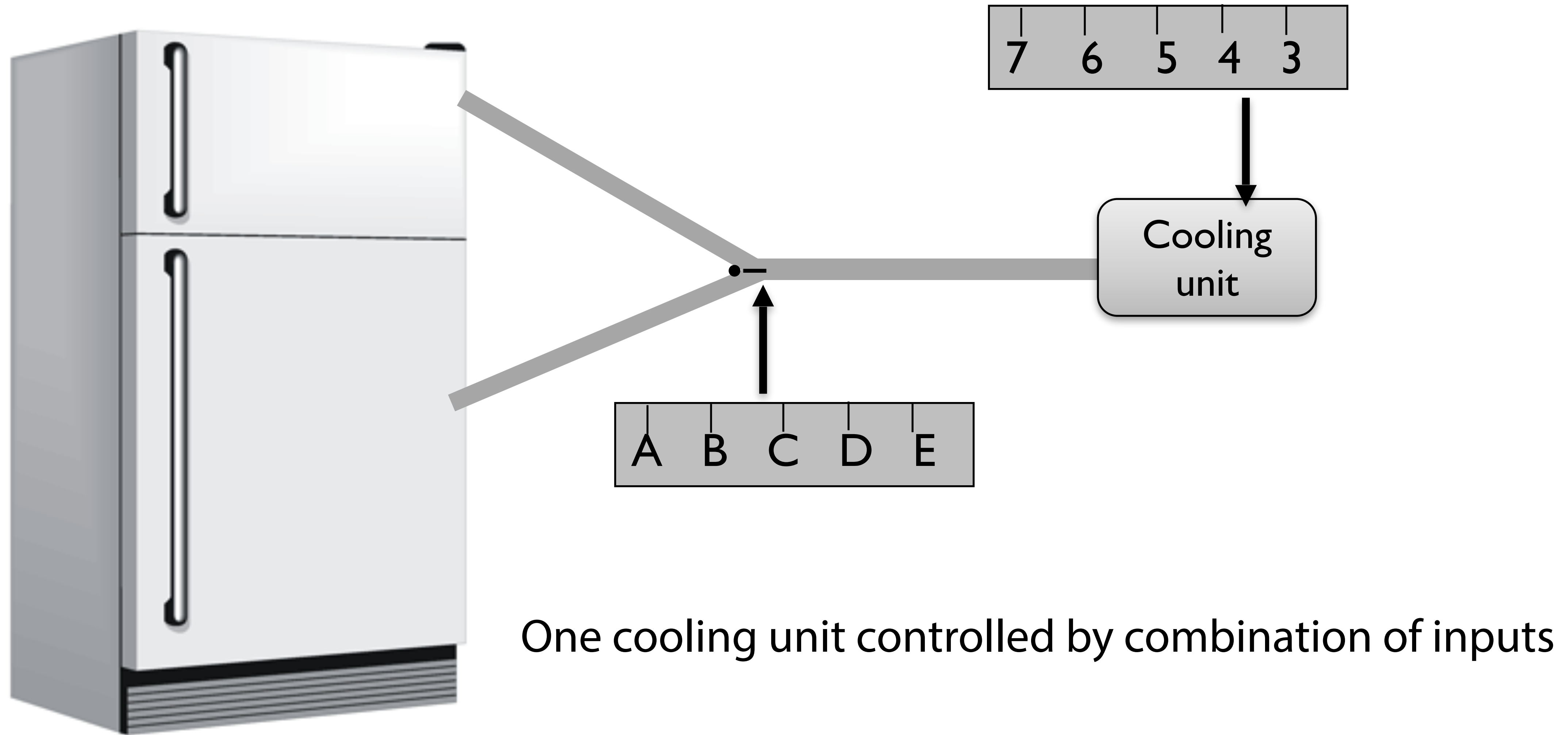
A	B	C	D	E	7	6	5	4	3
---	---	---	---	---	---	---	---	---	---

Freezer

Fresh Food



Independently controlled cooling units



One cooling unit controlled by combination of inputs

A **good conceptual model** let's users **predict** how **input controls** affect the **output**

When the **conceptual model** is **not predictive**, users resort to **trial-and-error**

It is **our job** as AI tool builders to provide interfaces that **let users build predictive conceptual models**

Gulf of Envisioning

[Subramonyam et al. 2024]

Capability gap: prompts don't know how to convert intentions into actions the AI can perform, because AI capabilities and actions are unclear

Gulf of Execution
semantic distance

Instruction gap: prompts don't always know how to state in natural language what they want the AI to do because language is ambiguous

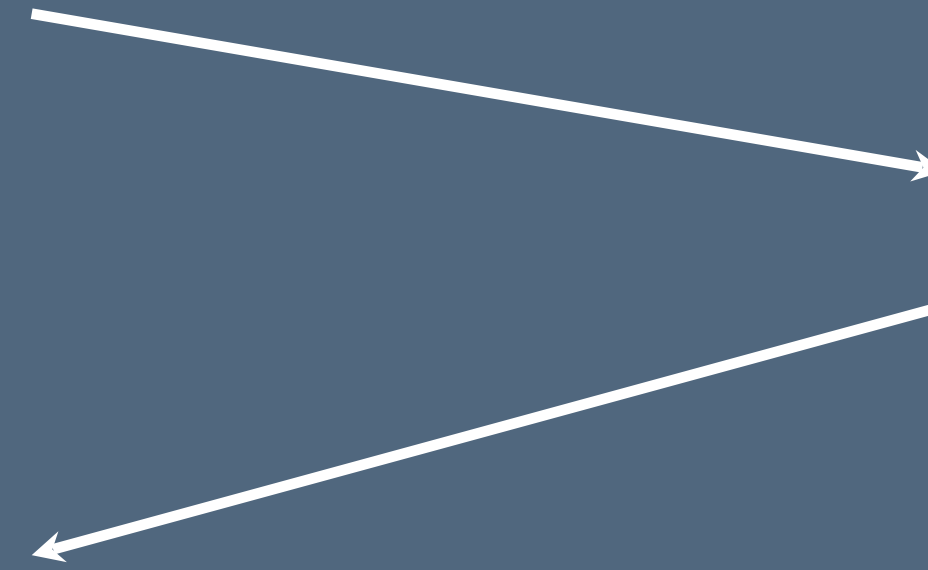
Gulf of Execution
articulatory distance

Intentionality gap: prompts don't always think about how to evaluate whether the results really meet their needs

Gulf of Evaluation
semantic distance

YOU READ THIS

Picture of a cool, young Computer Science Professor named Maneesh Agrawala

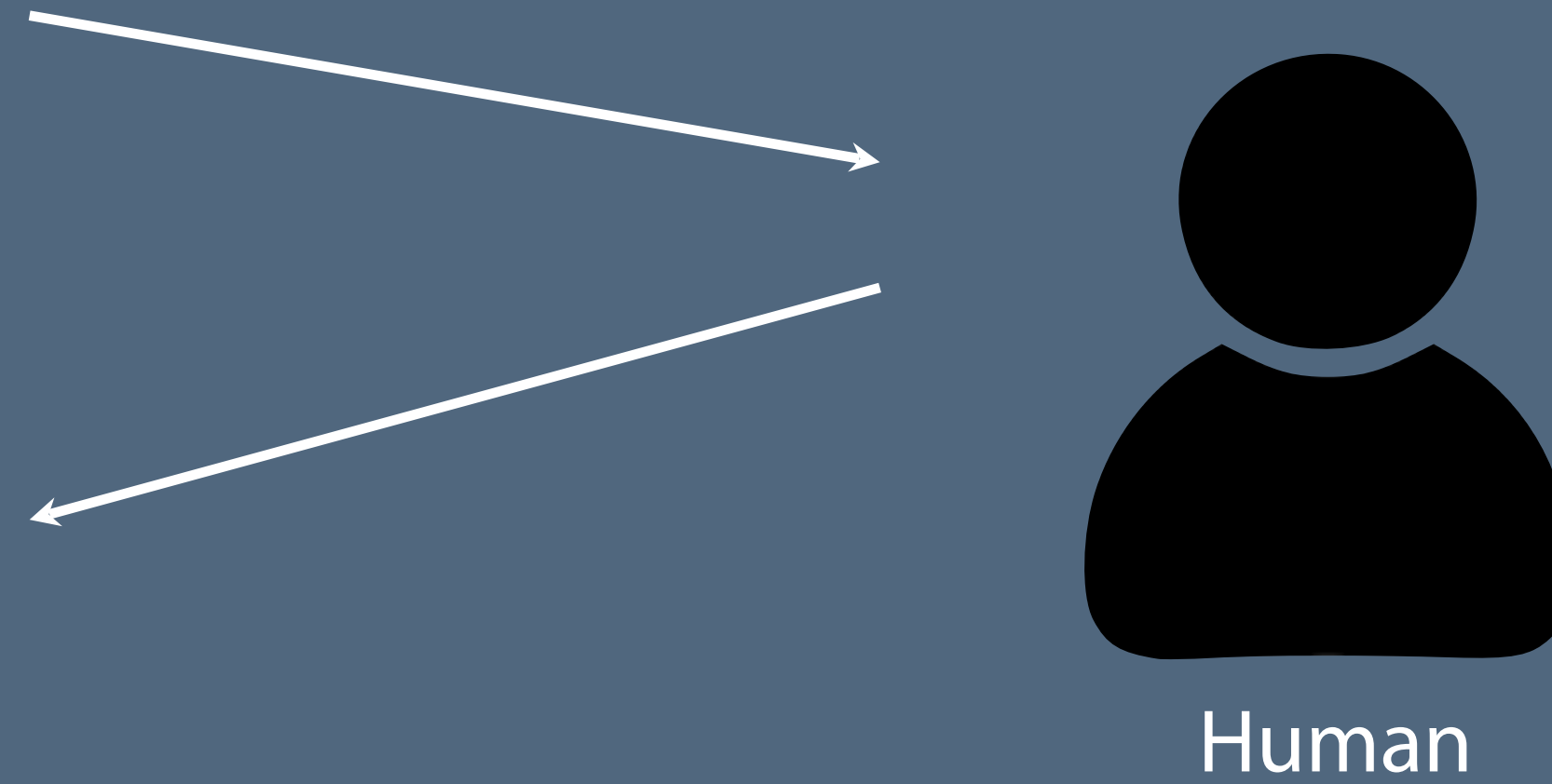


DALL-E2

AI black boxes are terrible interfaces

- Does “cool” imply a sportscoat?
- Does “picture” generate a photograph?
- Cannot predict how input prompt affects output image

Picture of a cool, young Computer Science Professor named Maneesh Agrawala



Humans
AI black boxes are terrible interfaces

- Does “cool” imply a sportscoat?
- Does “picture” generate a photograph?
- Cannot predict how input prompt affects output image

Interacting with a Human



Picture of a cool, young
Computer Science
Professor named
Maneesh Agrawala



Should I make him cool
by having him wear a
sportscoat or a hoodie?

Maybe something in
between

Do you want a
photograph or an
illustration?

illustration?



Interacting with a Human



Human

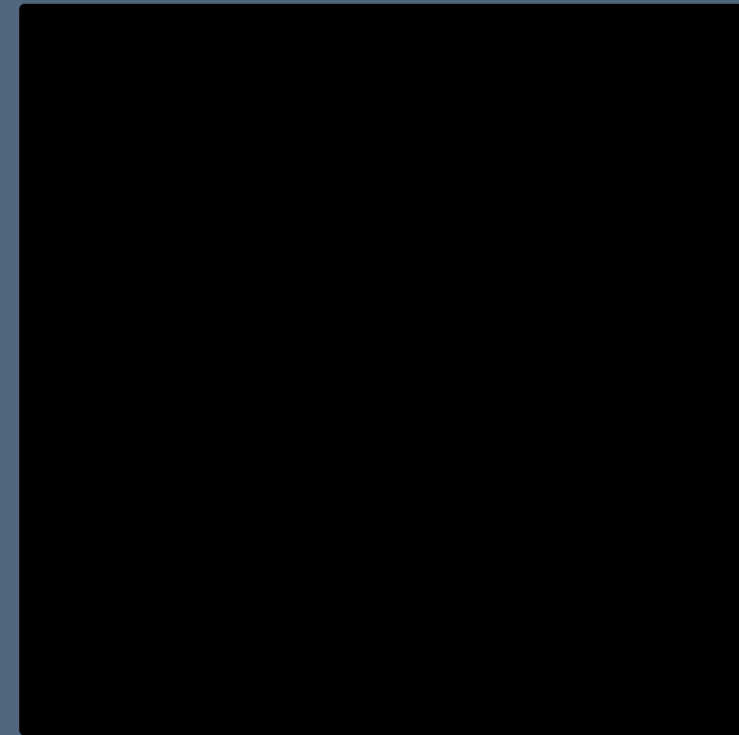
Conversational interaction

- Builds *common ground/shared semantics*
- Can *repair* to fix ambiguity/misunderstanding

Conceptual model based on model of self

But, some iteration still required

Interacting with a Black Box AI



AI

No conversation: each prompt generates new output

- **No** build up of *common ground/shared semantics*
- **No** *repair* to fix ambiguity/misunderstanding

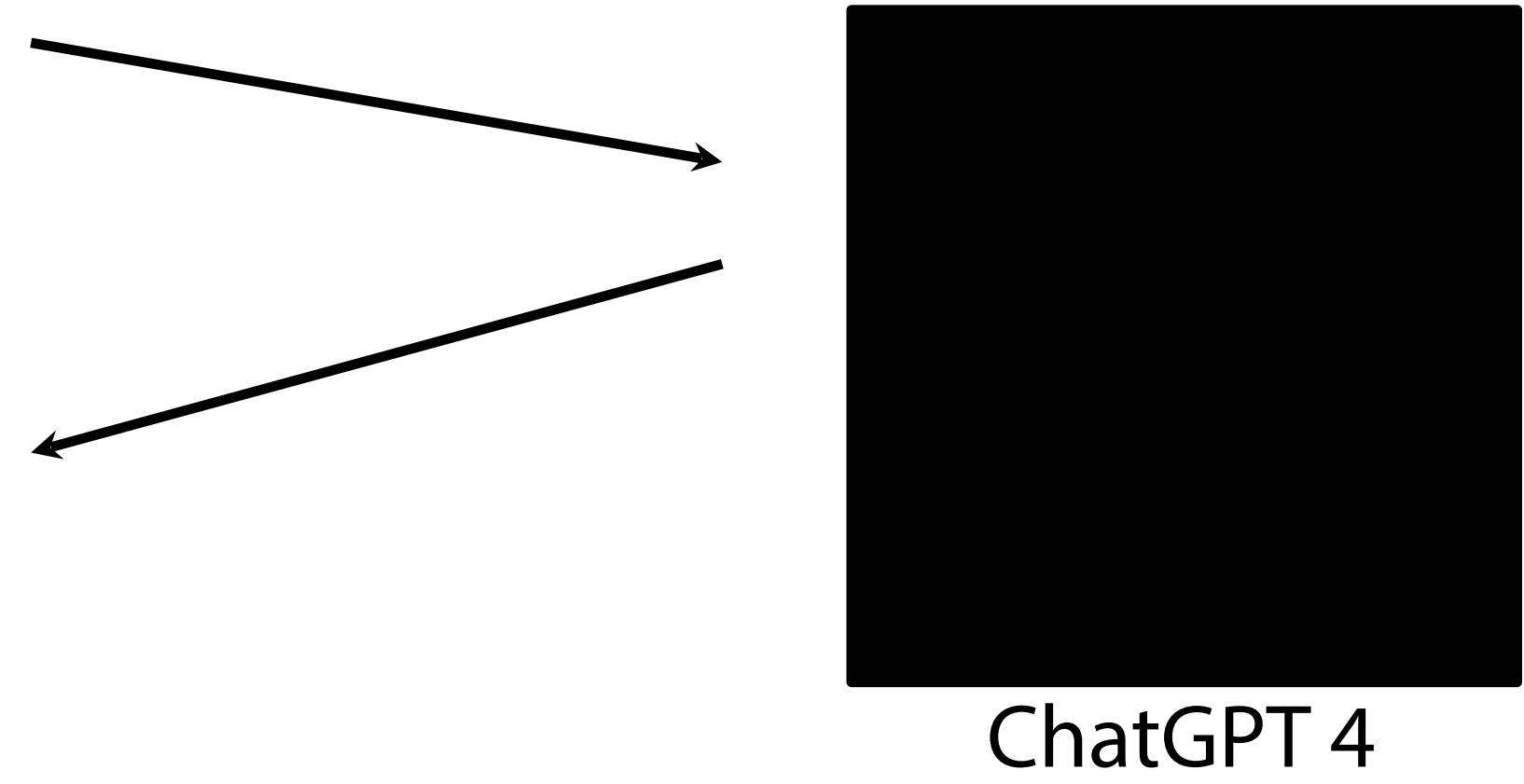
Conceptual model either **non-existent** or **incorrect** (based on self)

Lots of trial-and-error

Towards Conversational AI Interfaces

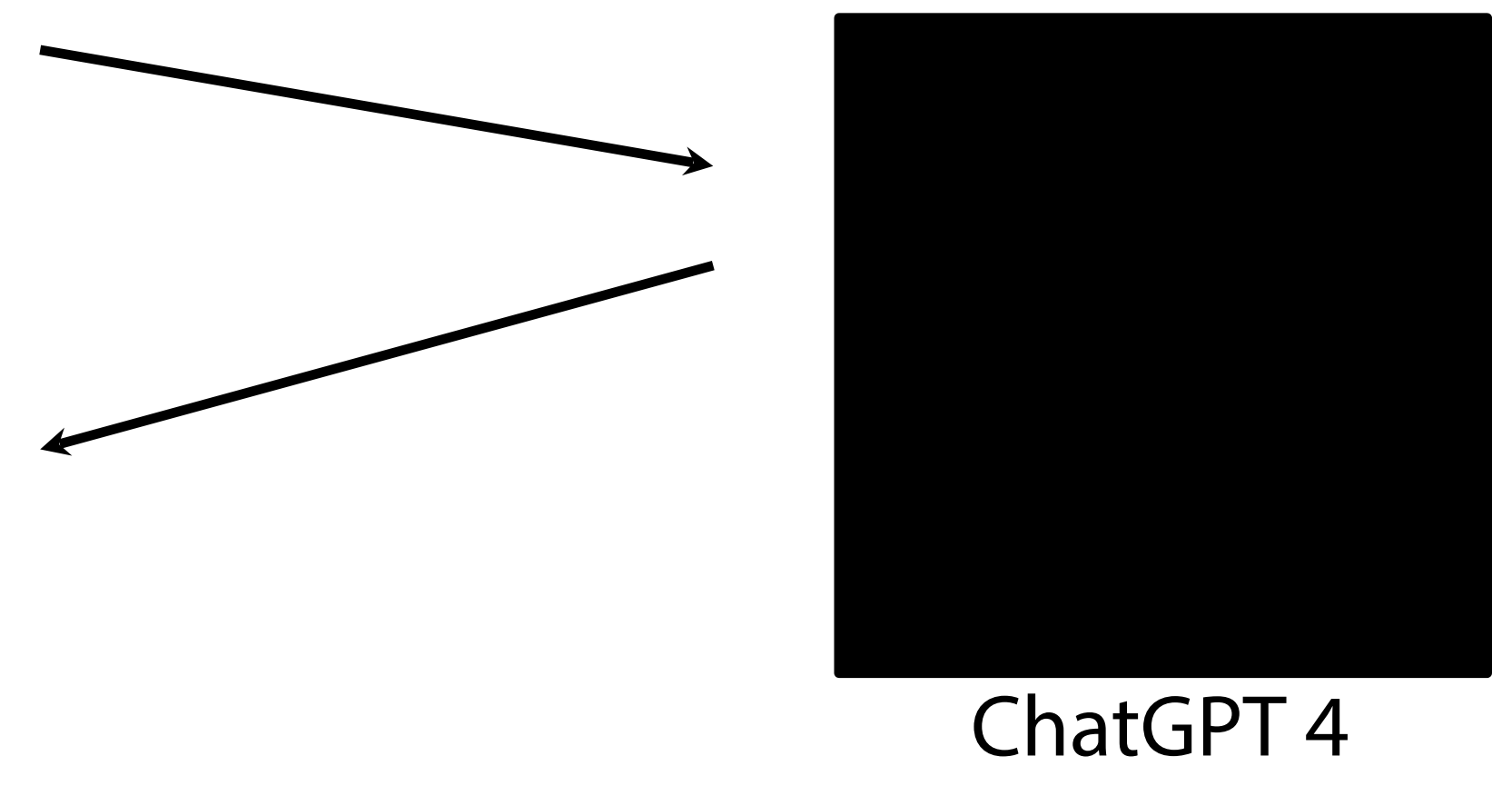


Who is the smartest professor?



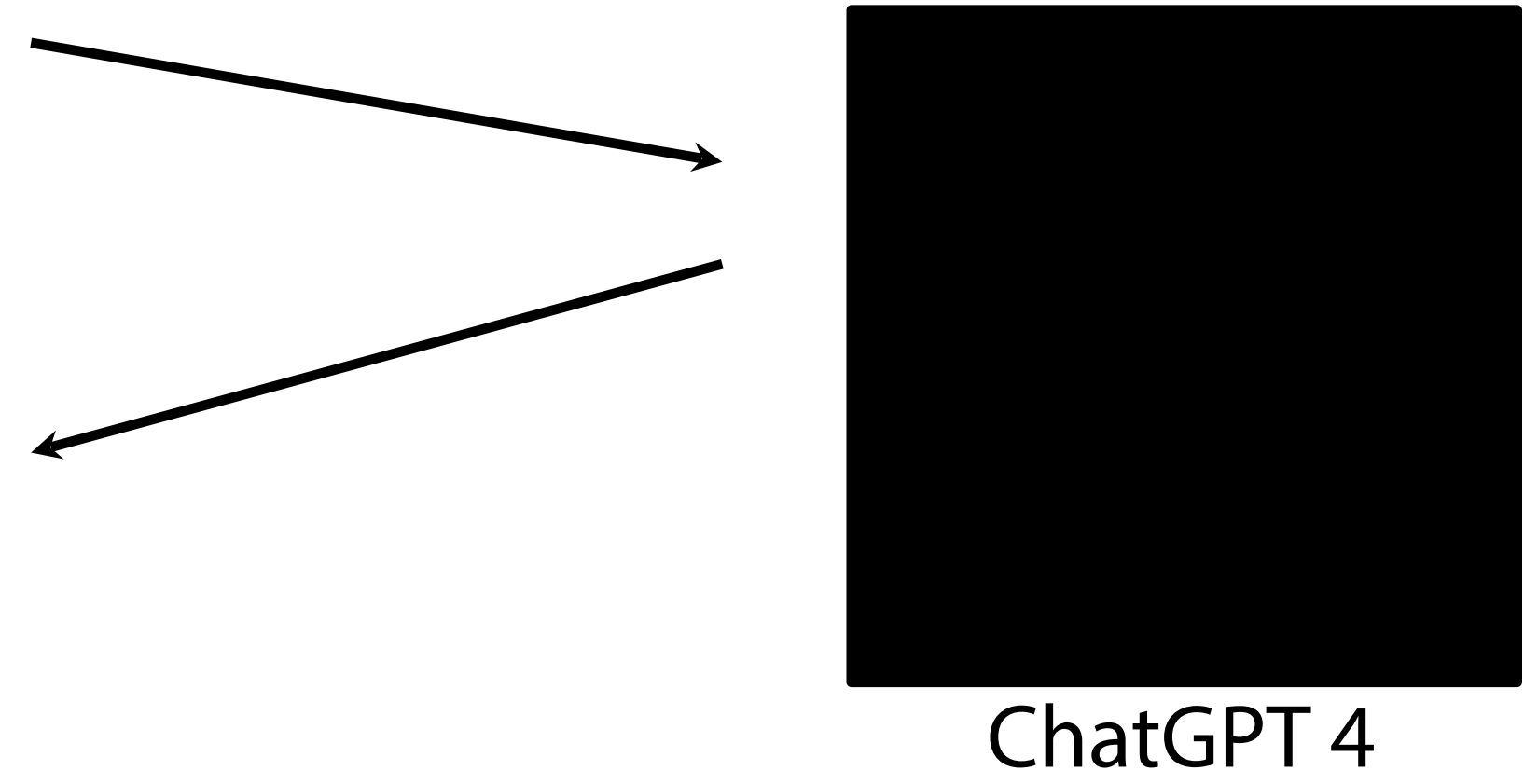
MA

By smartest I mean smartest in HCI



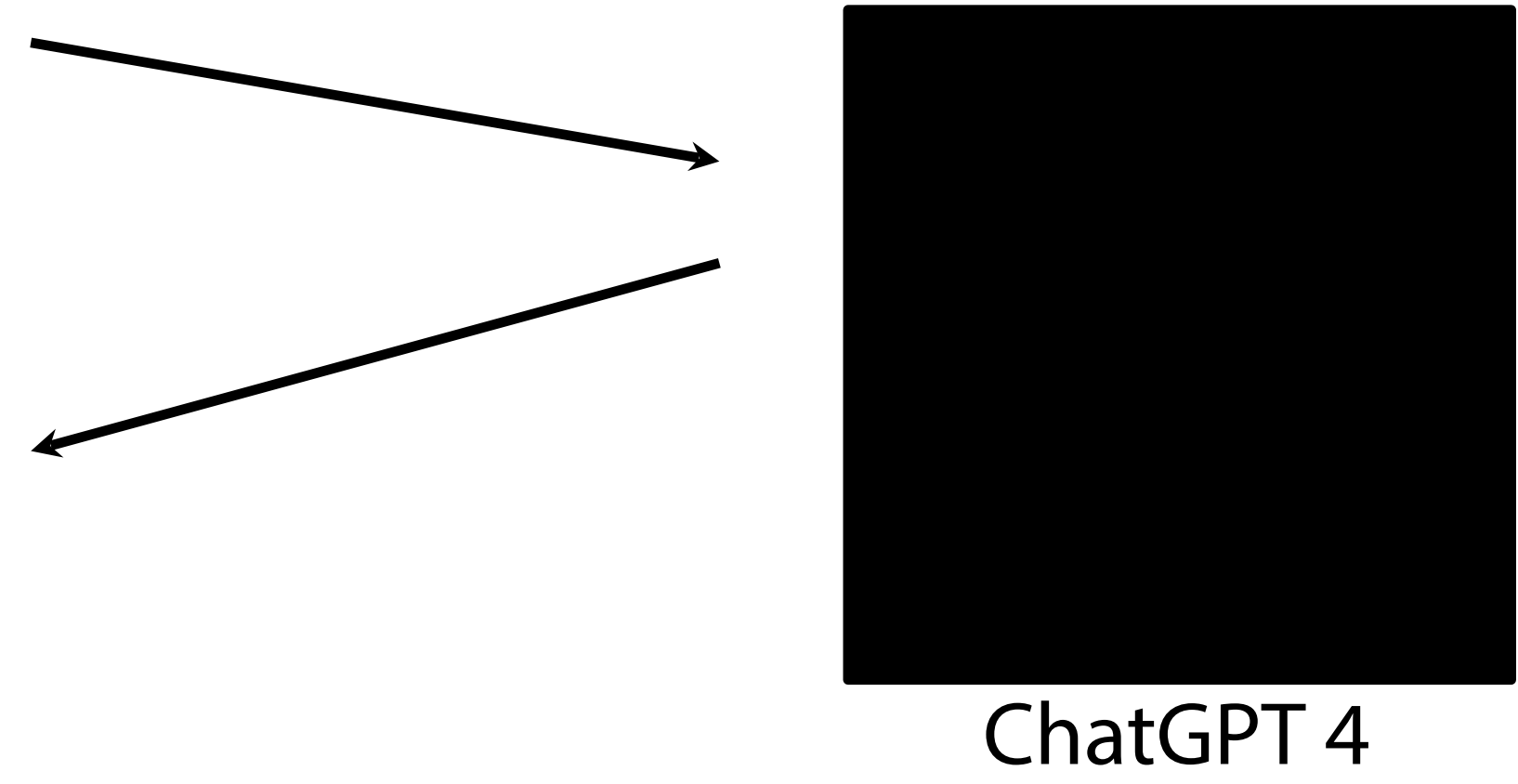
MA

And who is smartest in the Bay Area?



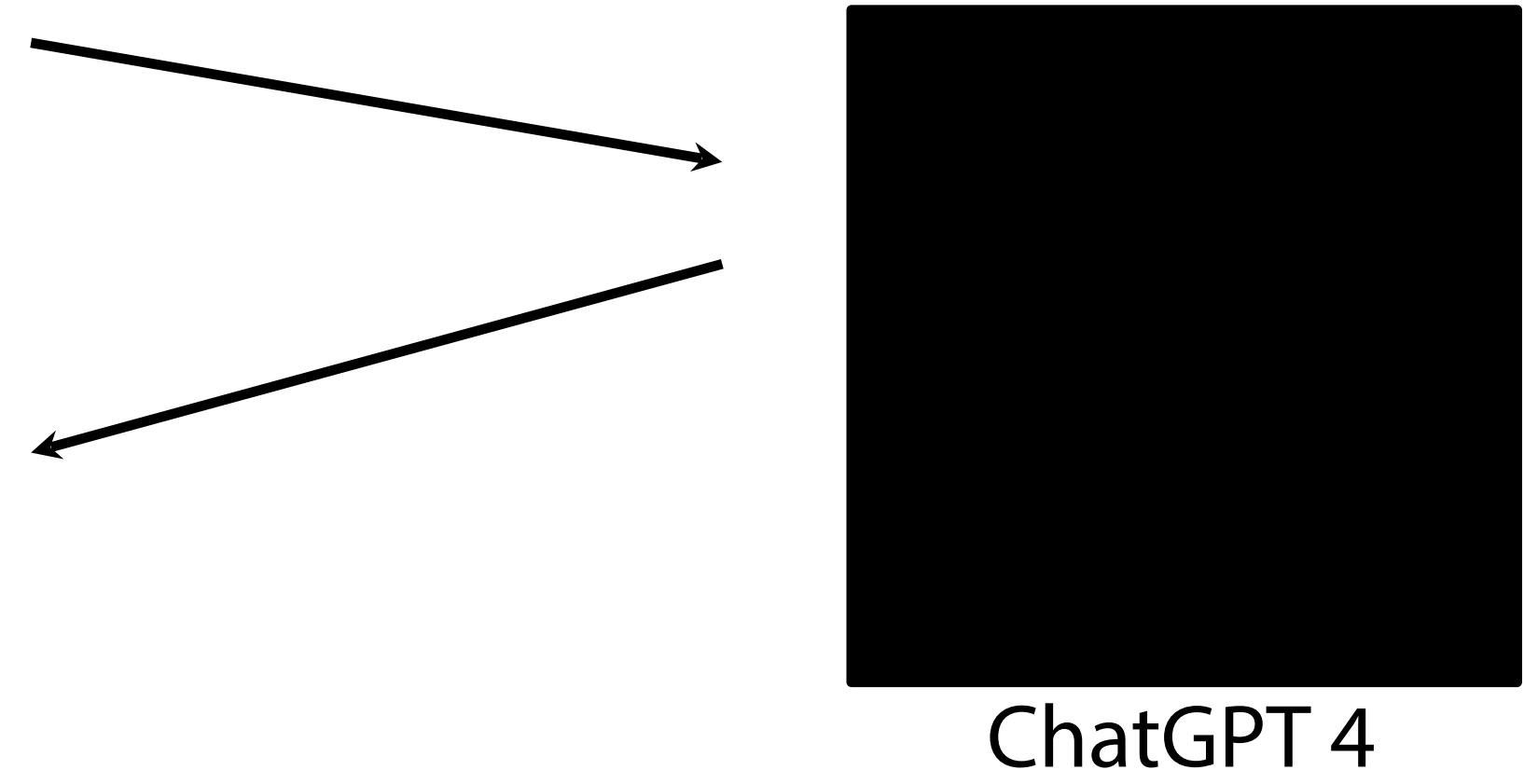
MA

And who is smartest **in this way** in the Bay Area?



MA

And who is smartest **in this way at Stanford?**



Conversational Interactions with ChatGPT

Support for turn-taking and context enables some common ground

- AI and human **can refer to concepts** from earlier exchanges **to refine them**
- But **refinement is one-sided**. AI doesn't ask for refinement – human adds it
- **AI model does not immediately learn** (or update weights) from the conversation.
- **Grounding seems shallow** (unclear what ChatGPT knows and doesn't know/remember)

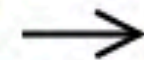
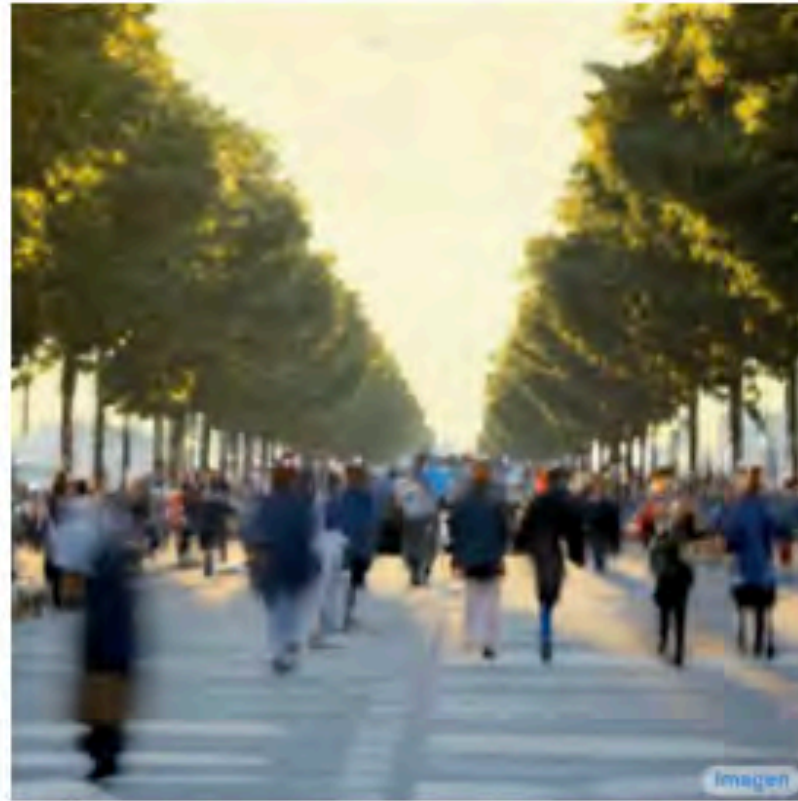
Establishing Common Ground



Input images

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation [Ruiz 2022]

Repair



“The boulevards are crowded today.”



“Children drawing of a castle next to a river.”



“Photo of a cat riding on a ~~bicycle~~ car.”



“a cake with decorations.”

jelly beans

Dealing with Ambiguity of Natural Language





Prompt: full body, walking pose, slow motion, female spiderman wearing full body (light silver armour:1.2), (insanely detailed, bloom:1.5), (highest quality, Alessandro Casagrande, Greg Rutkowski, Sally Mann, concept art, 4k), (analog:1.2), (high sharpness), (detailed pupils:1.1), (painting:1.1), (digital painting:1.1), detailed face and eyes, Masterpiece, best quality, (highly detailed photo:1.1), 8k, photorealistic, (long blonde Hair, ponytail haircut, ecstatic:1.1), (young woman:1.1), By jeremy mann, by sandra chevrier, by maciej kuciara, sharp, (perfect body:1.1), realistic, real shadow, 3d, (cold background:1.2), (by Michelangelo)



Prompt: full body, walking pose, slow motion, female spiderman wearing full body (light silver armour:1.2), (insanely detailed, bloom:1.5), (highest quality, Alessandro Casagrande, Greg Rutkowski, Sally Mann, concept art, 4k), (analog:1.2), (high sharpness), (detailed pupils:1.1), (painting:1.1), (digital painting:1.1), detailed face and eyes, Masterpiece, best quality, (highly detailed photo:1.1), 8k, photorealistic, (long blonde Hair, ponytail haircut, ecstatic:1.1), (young woman:1.1), By jeremy mann, by sandra chevrier, by maciej kuciara, sharp, (perfect body:1.1), realistic, real shadow, 3d, (cold background:1.2), (by Michelangelo)

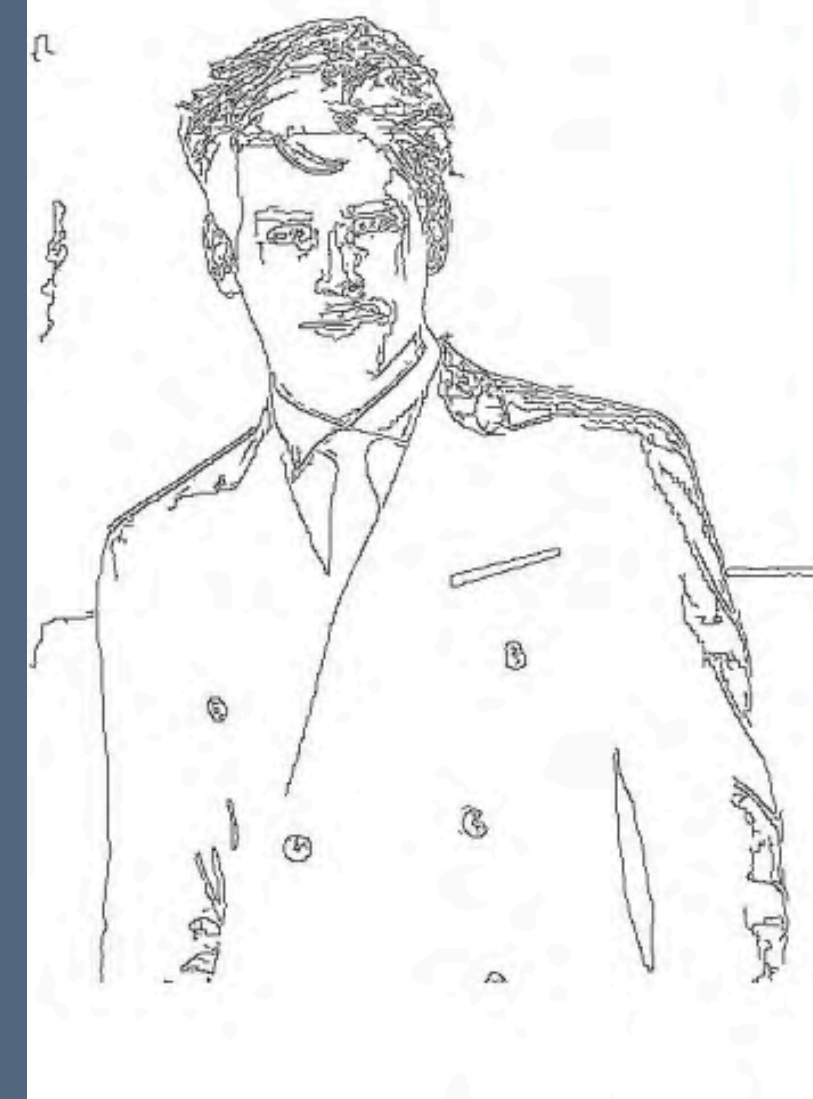
Problem: prompt provides little spatial control over composition and pose



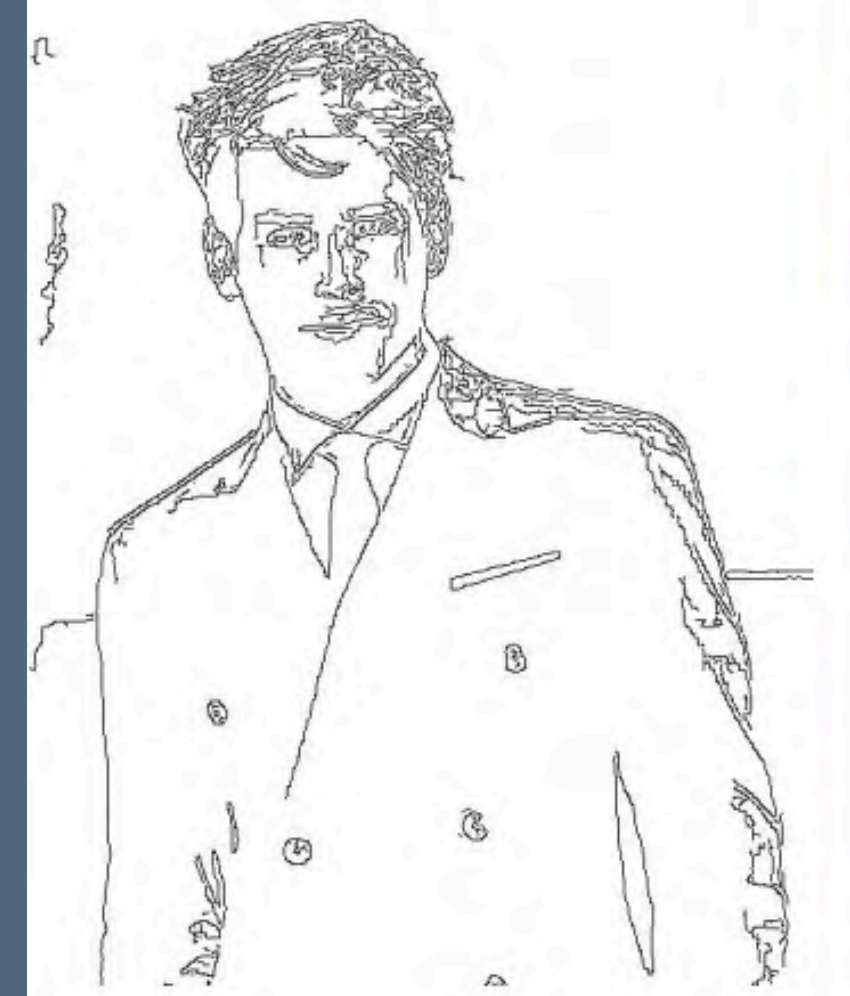
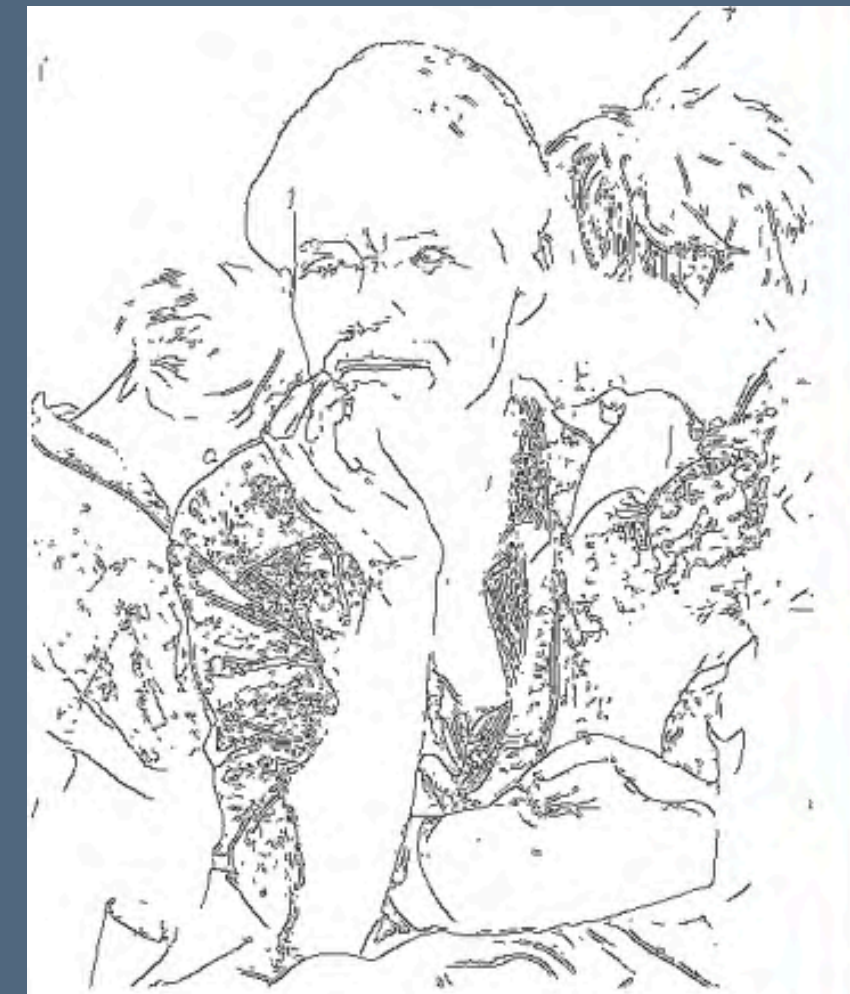
Prompt: full body, ~~walking~~ **swinging** pose, slow motion, female spiderman wearing full body (light silver armour: 1.2), (insanely detailed, bloom: 1.5), (highest quality, Alessandro Casagrande, Greg Rutkowski, Sally Mann, concept art, 4k), (analog: 1.2), (high sharpness), (detailed pupils: 1.1), (painting: 1.1), (digital painting: 1.1), detailed face and eyes, Masterpiece, best quality, (highly detailed photo: 1.1), 8k, photorealistic, (long blonde Hair, ponytail haircut, ecstatic: 1.1), (young woman: 1.1), By jeremy mann, by sandra chevrier, by maciej kuciara, sharp, (perfect body: 1.1), realistic, real shadow, 3d, (cold background: 1.2), (by Michelangelo)

Problem: prompt provides little spatial control over composition and pose

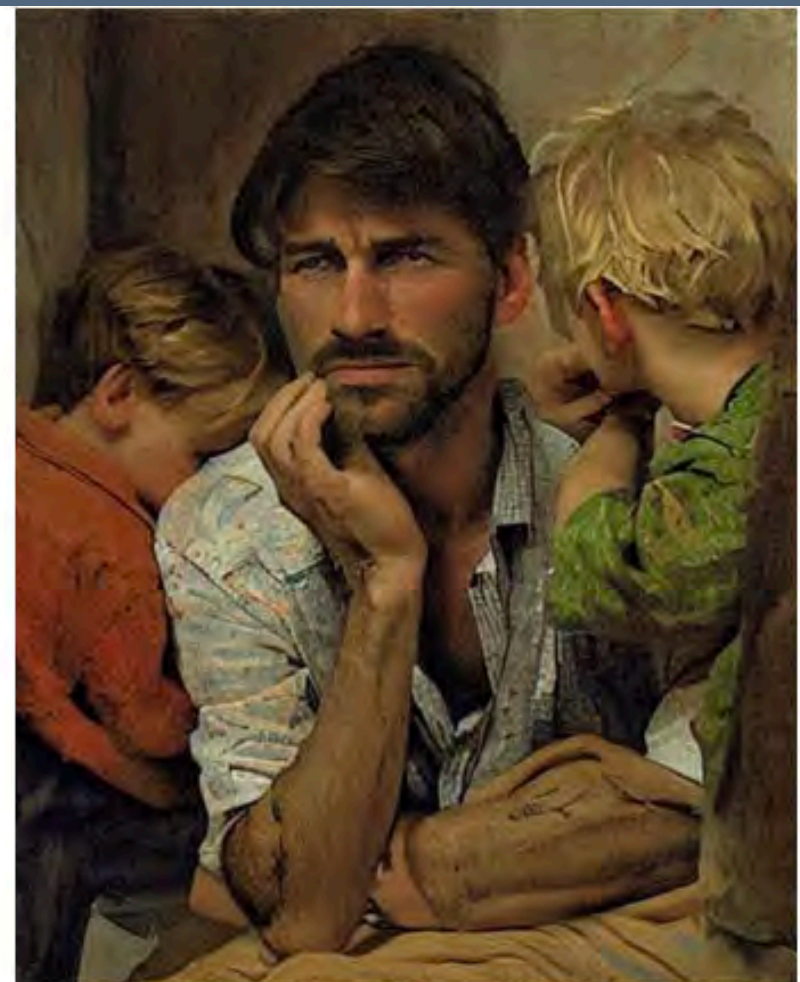
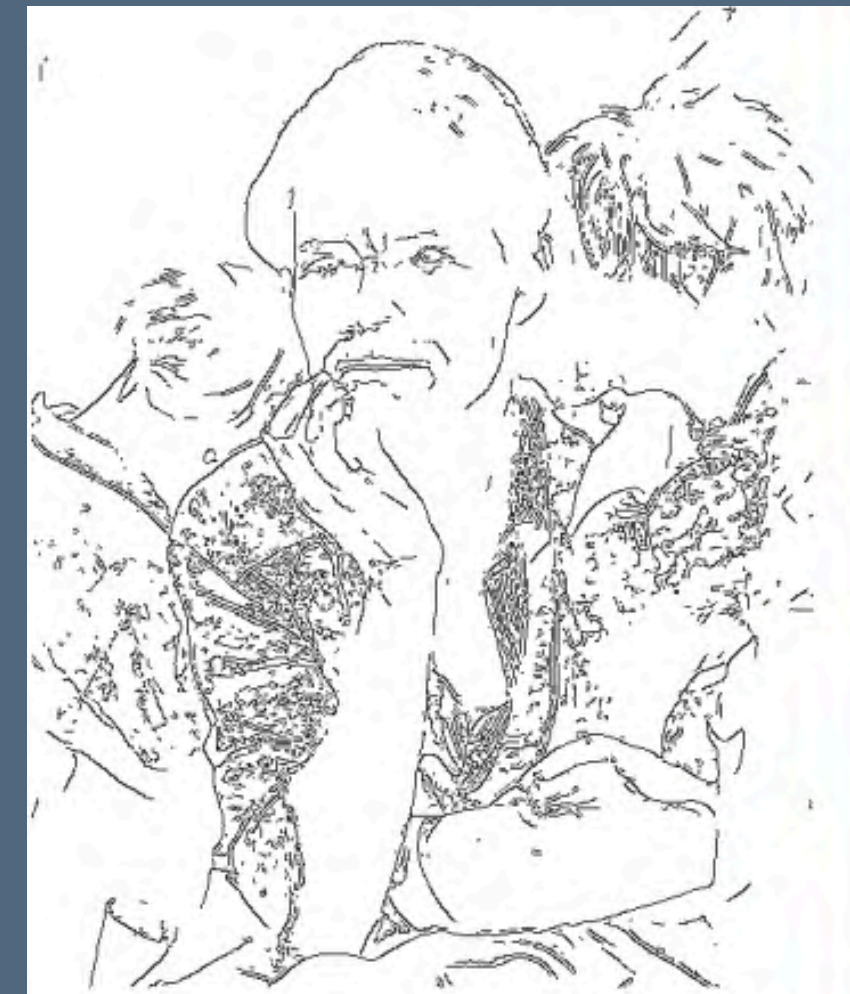
small changes to the prompt completely change the image composition



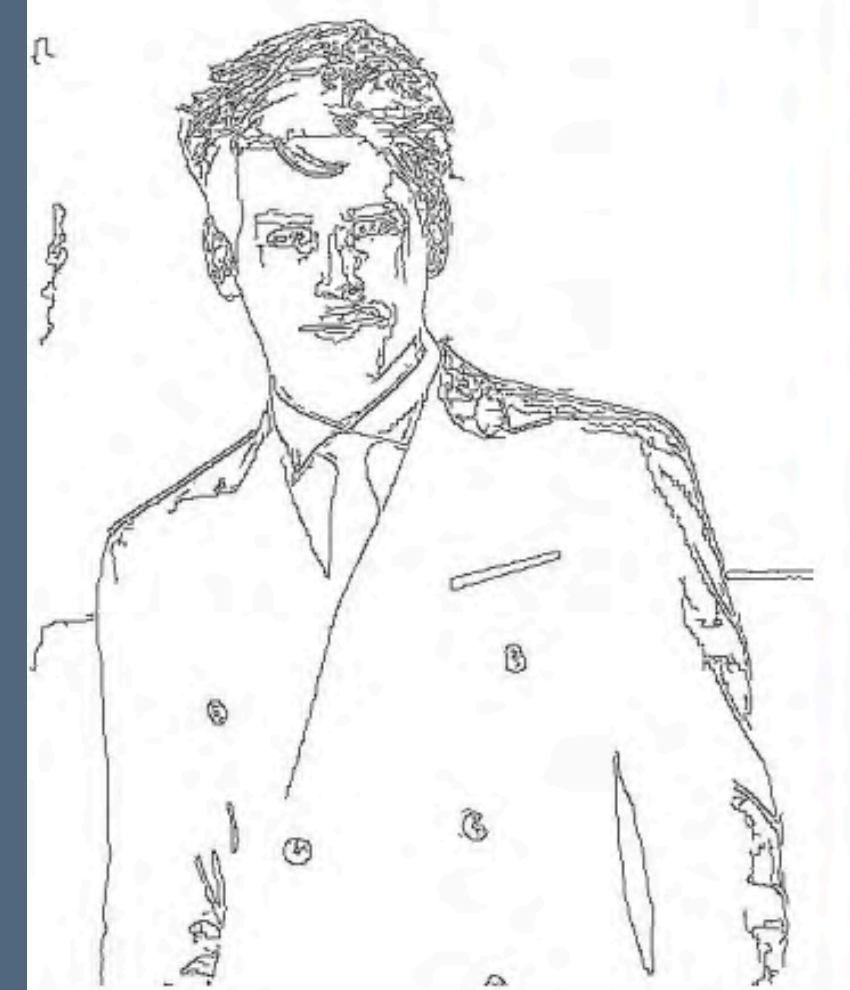
Adding Conditional Control to Text-to-Image Diffusion Models [Zhang 2023]



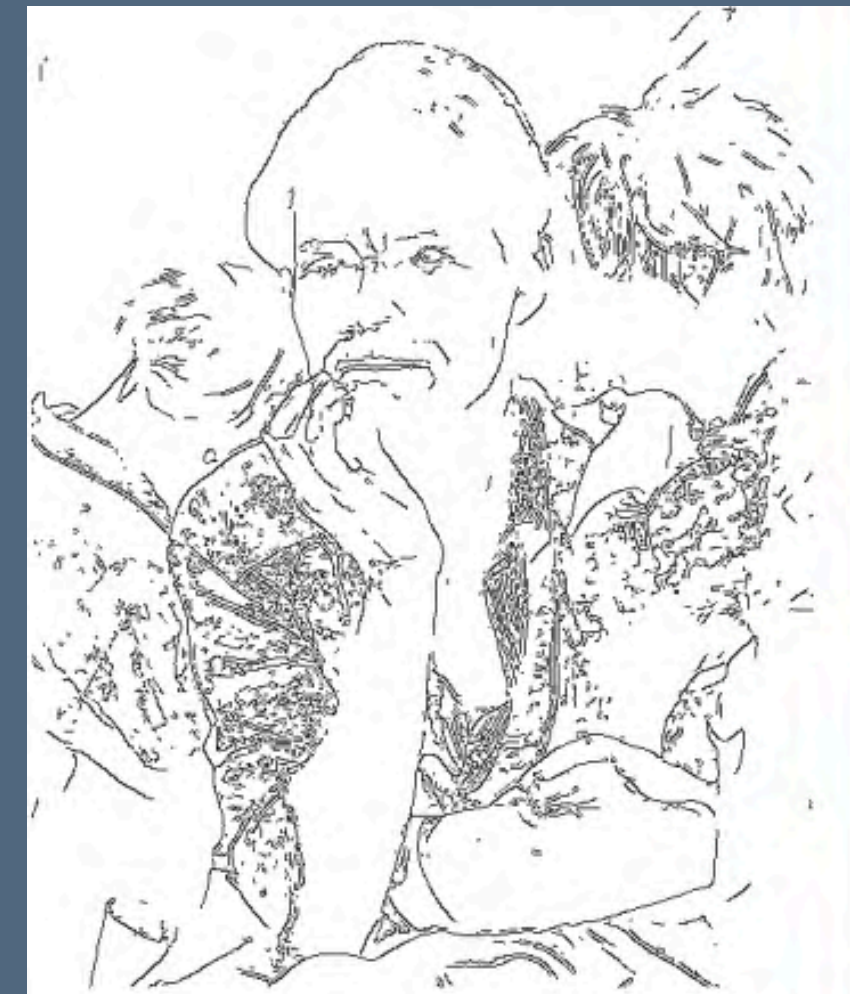
Adding Conditional Control to Text-to-Image Diffusion Models [Zhang 2023]



“a man with beard sitting with two children”

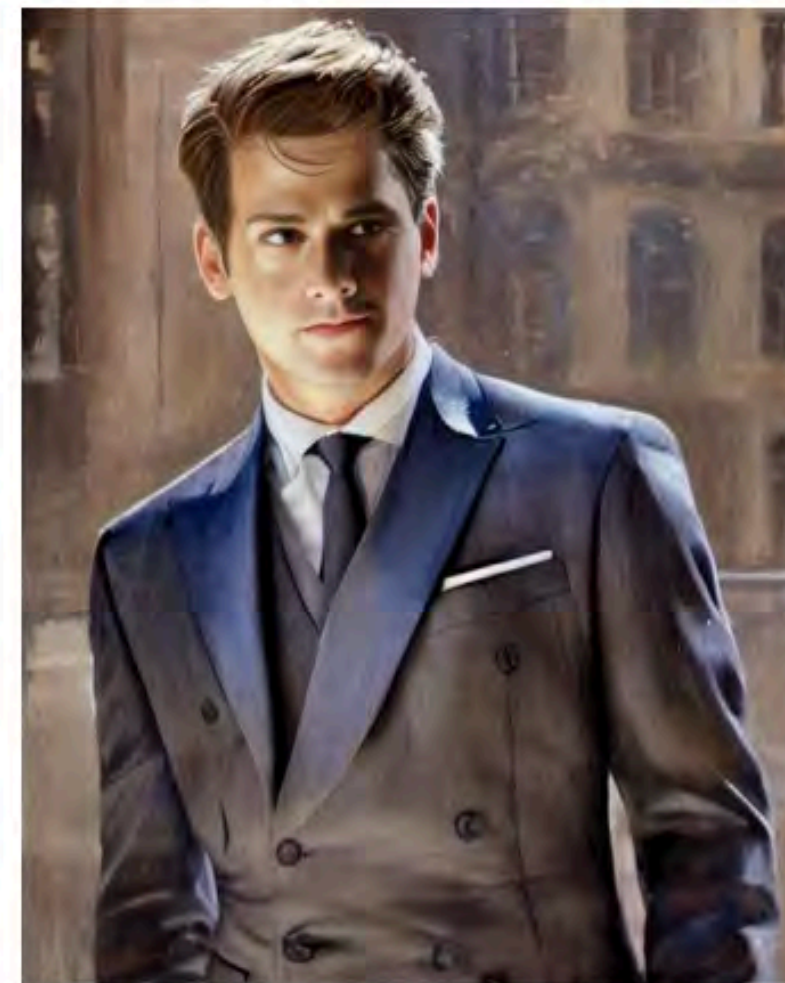
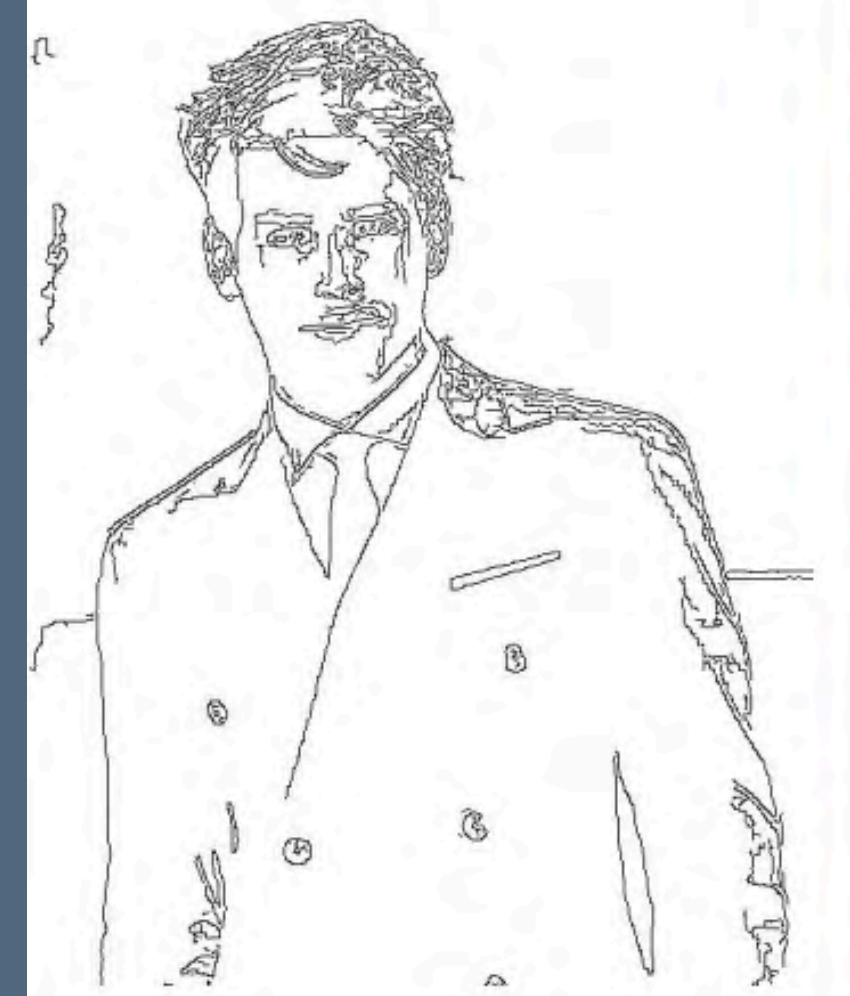


“a man in a suit and tie”



“a man with beard sitting with two children”

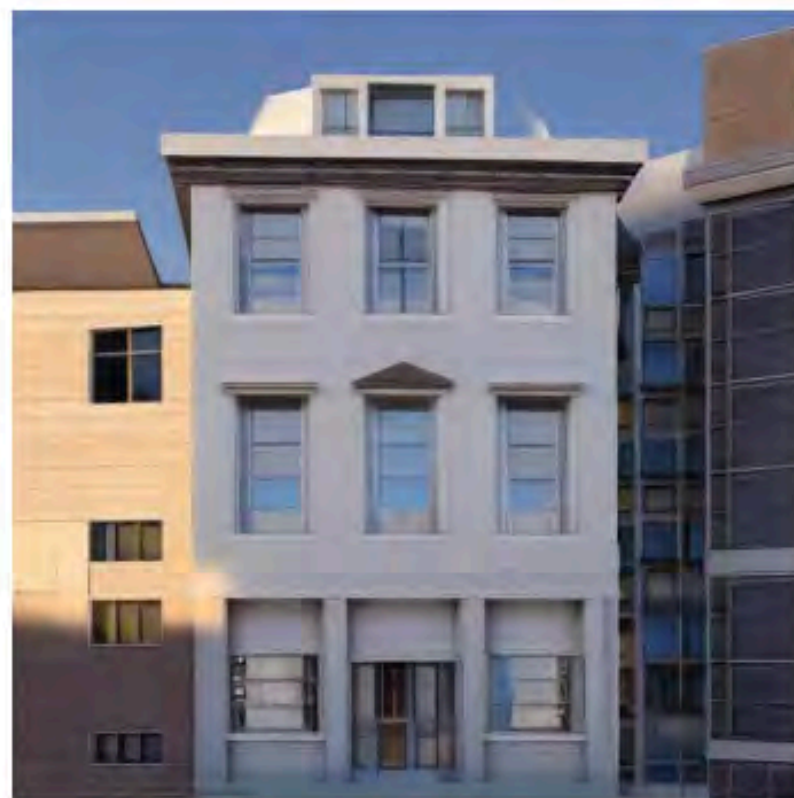
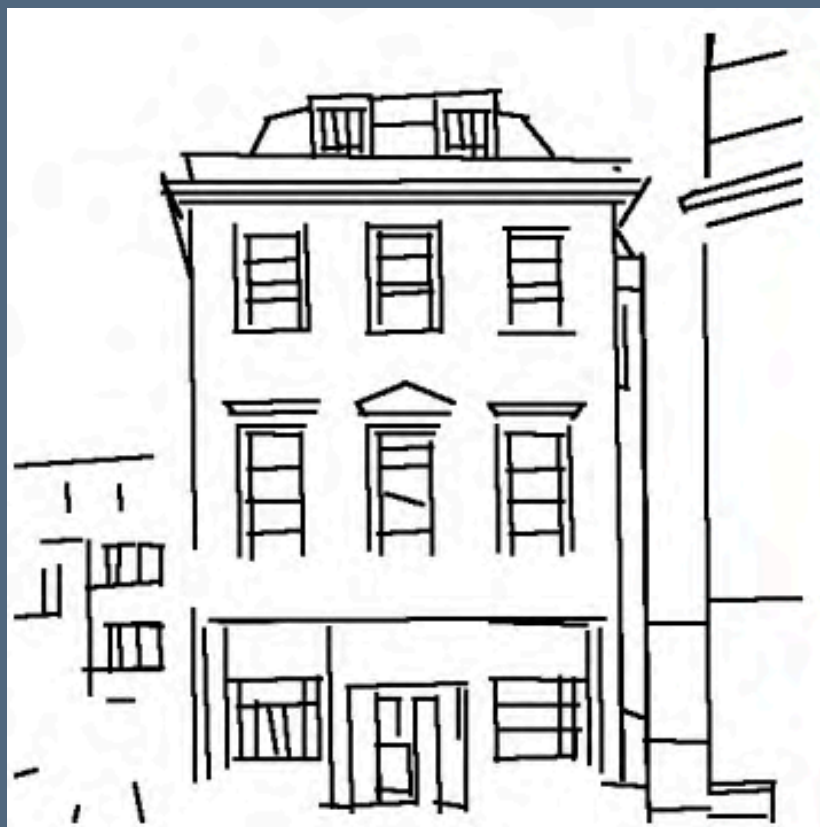
“mother and two boys in a room, masterpiece, artwork”



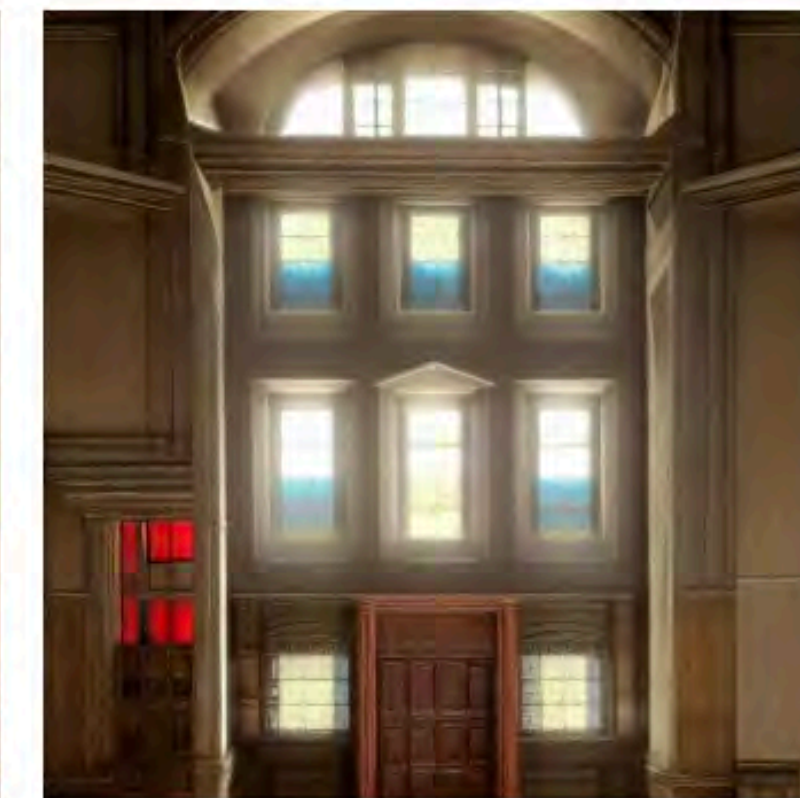
“a man in a suit and tie”

“a man in a white suit and tie”

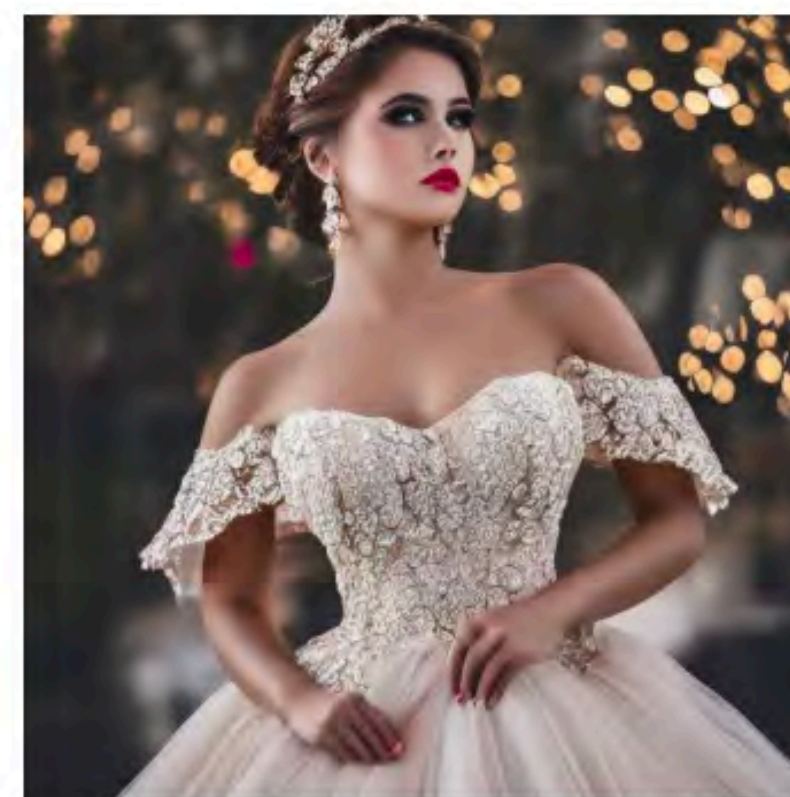
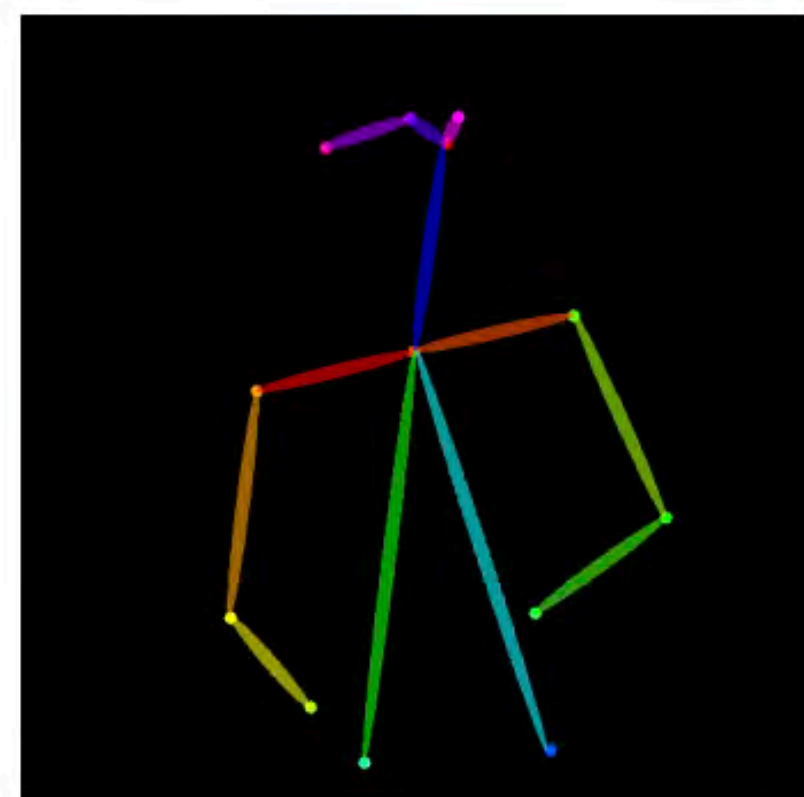
Dealing with Ambiguity of Spatial Language



“a building in a city street”



“inside a gorgeous 19th century church”

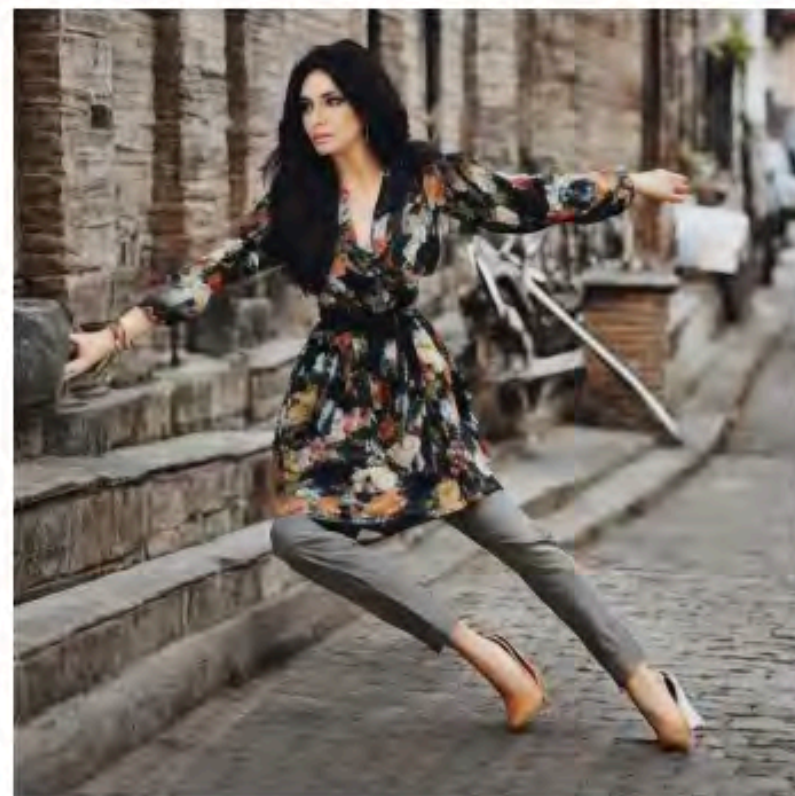
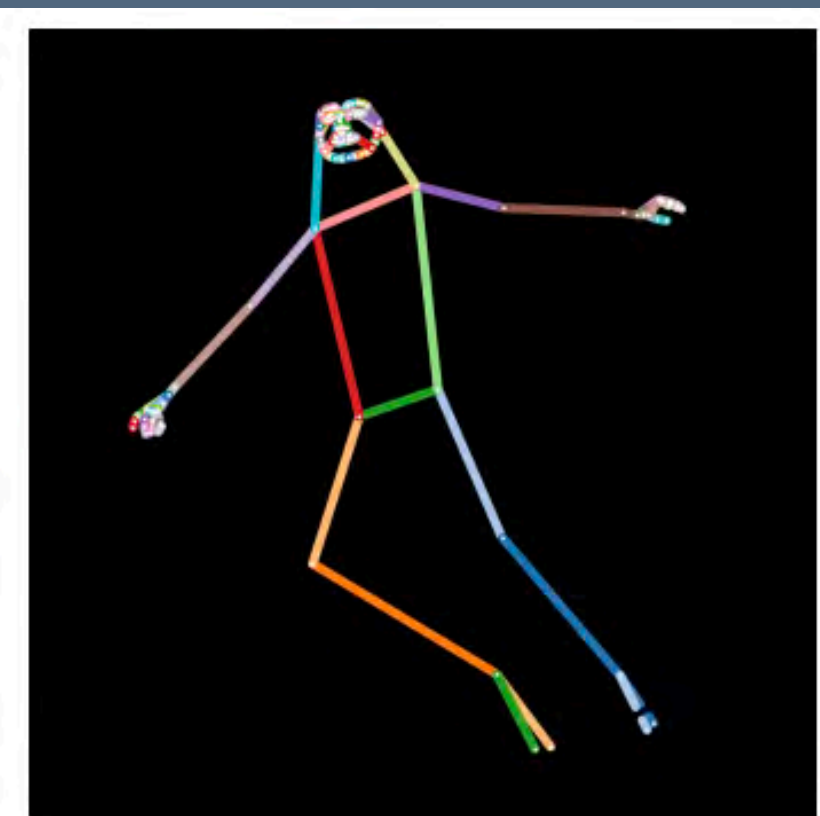


“chef in the kitchen”

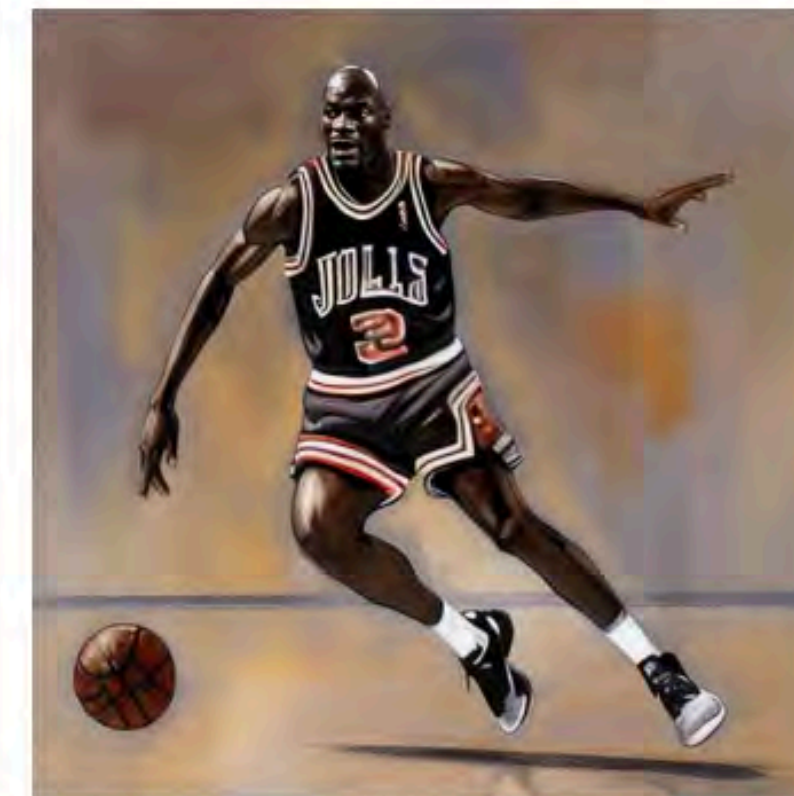
Idea: User provides *conditioning* image that puts spatially localized constraints on the output image

Adding Conditional Control to Text-to-Image Diffusion Models [Zhang 2023]

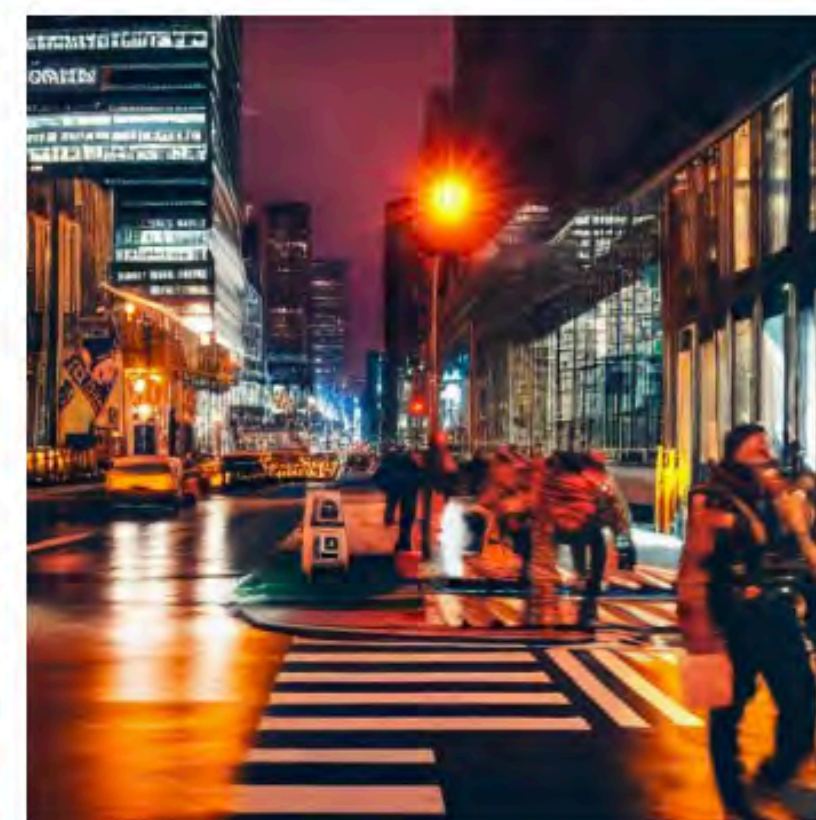
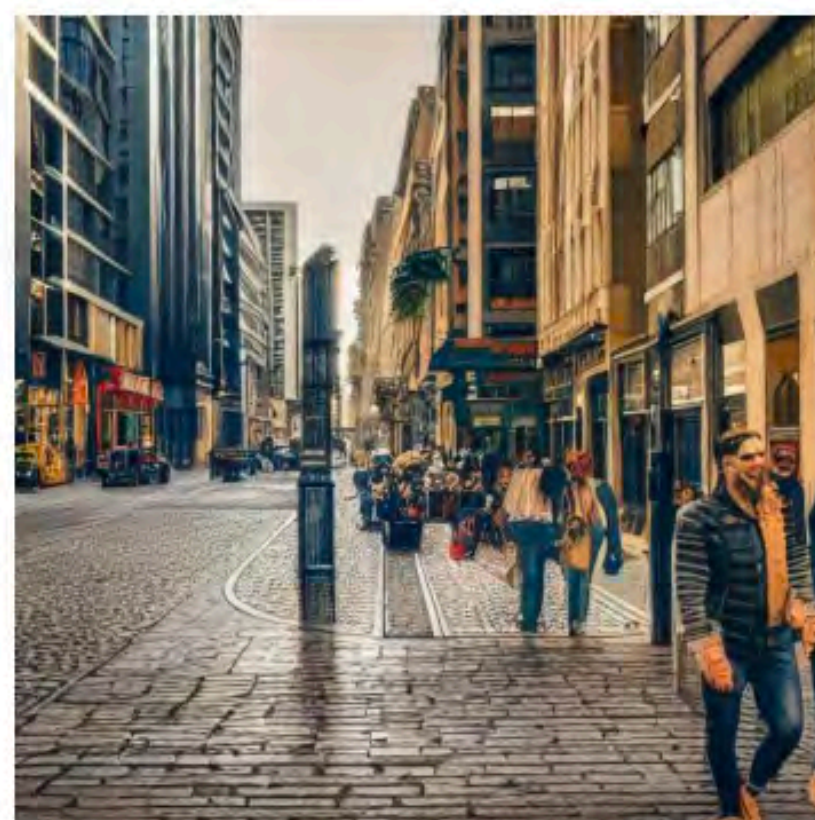
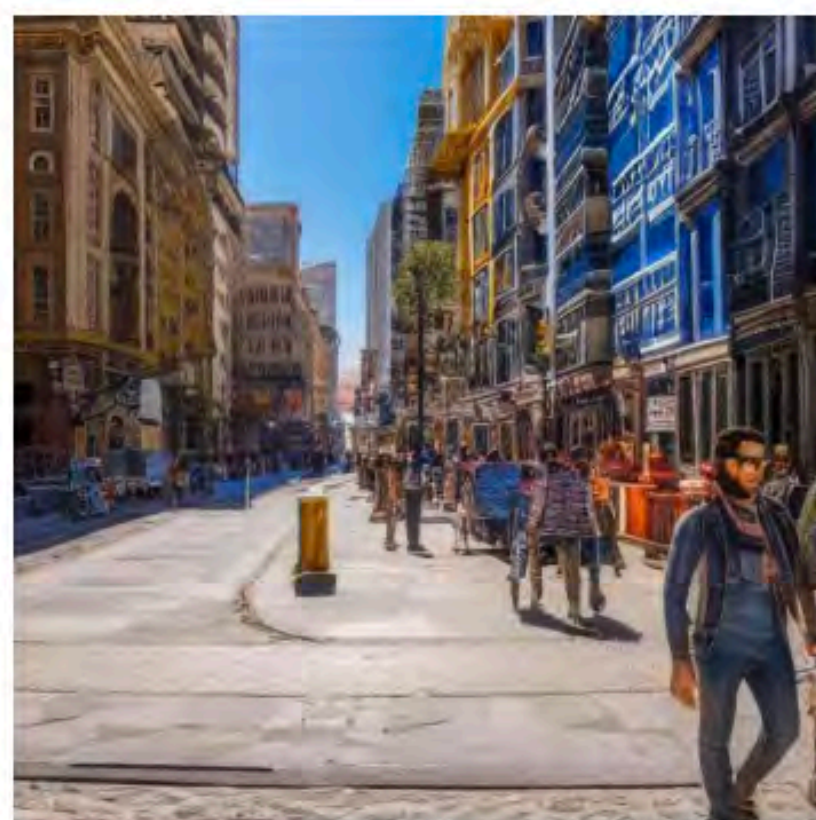
Dealing with Ambiguity of Spatial Language



“a woman dancing near a street corner”



“artwork of Michael Jordan playing basketball”



Idea: User provides *conditioning* image that puts spatially localized constraints on the output image

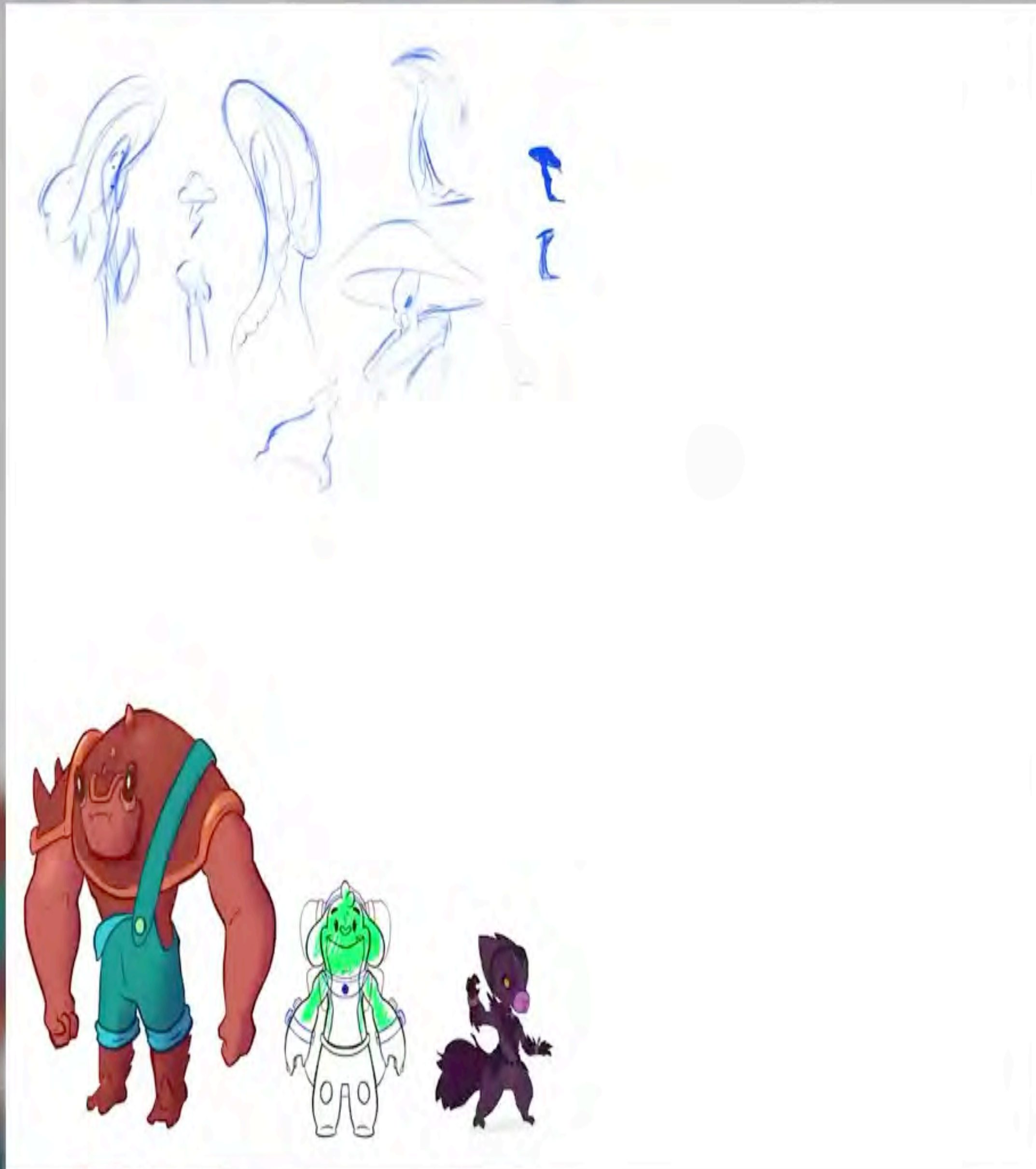
Adding Conditional Control to Text-to-Image Diffusion Models [Zhang 2023]

Iterative Refinement (not Iterative Trial-and-Error)

Incremental Actions

Maintain shared structure before
and after action

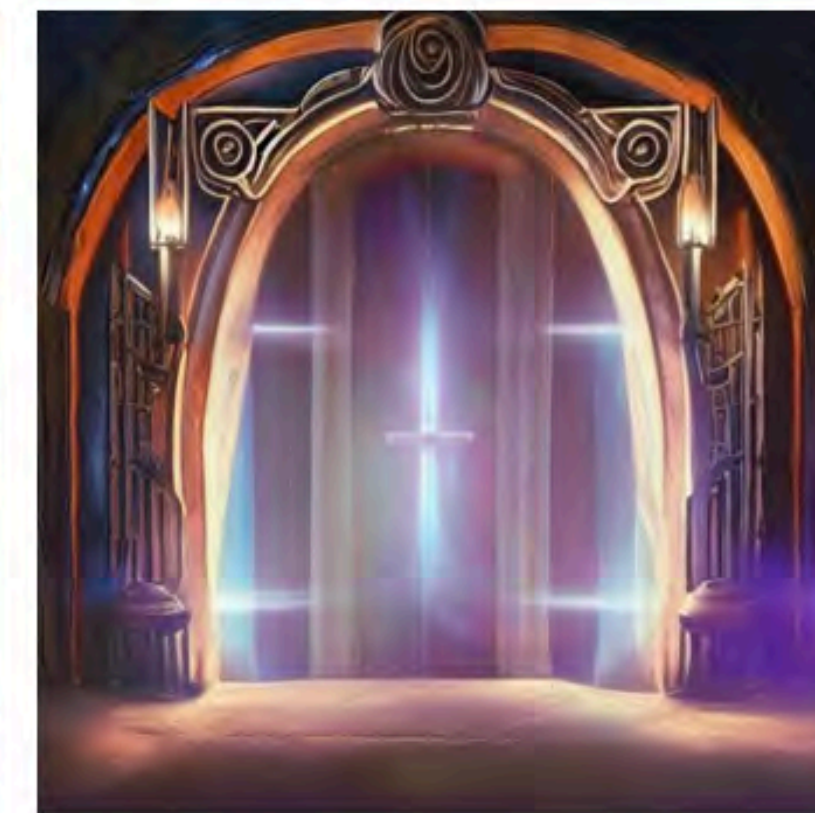
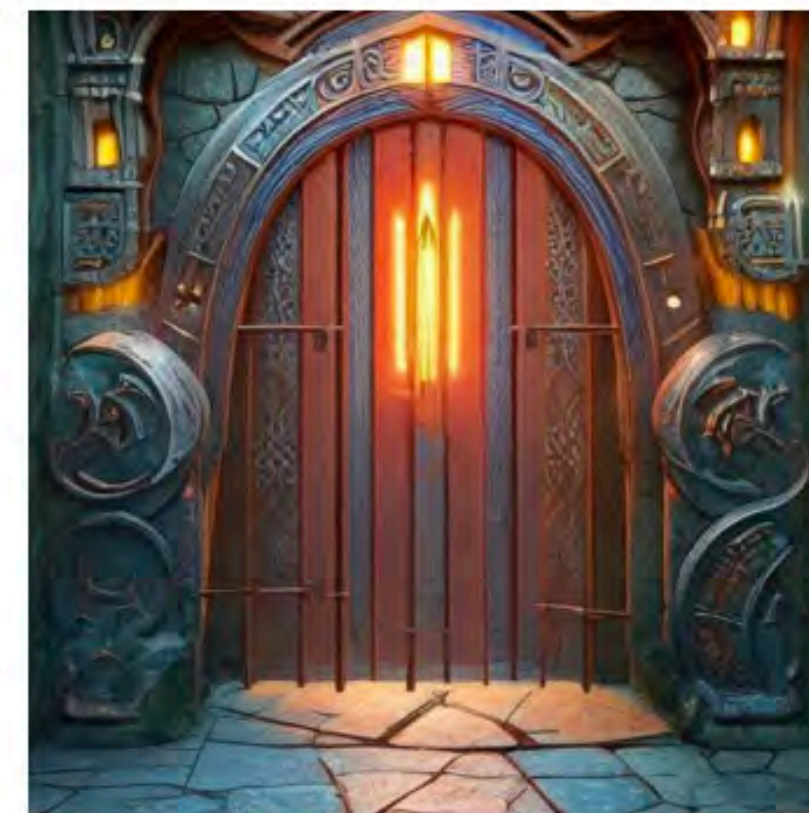
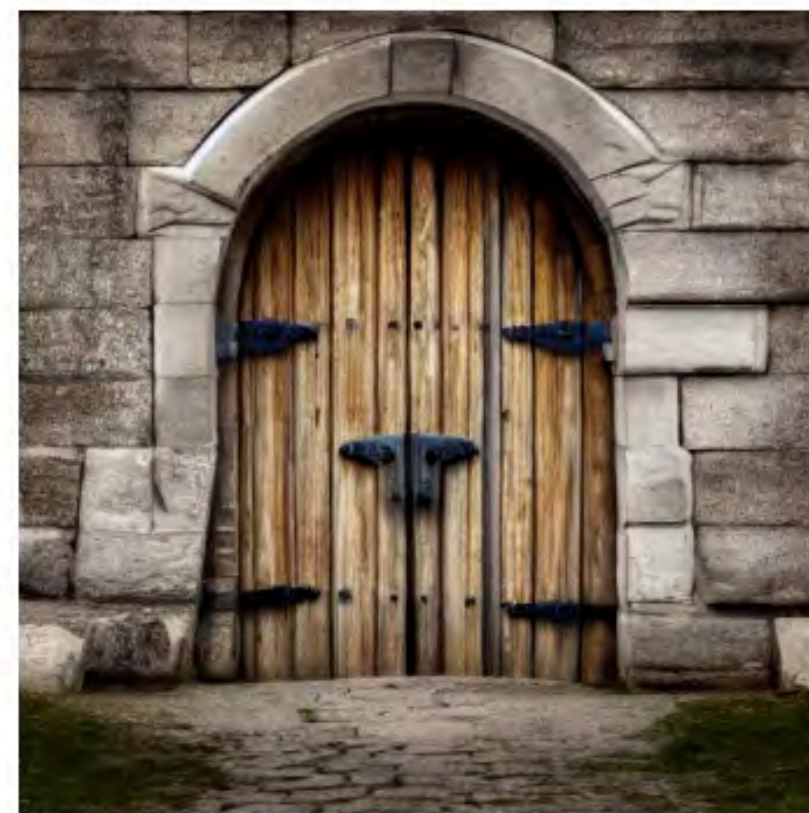
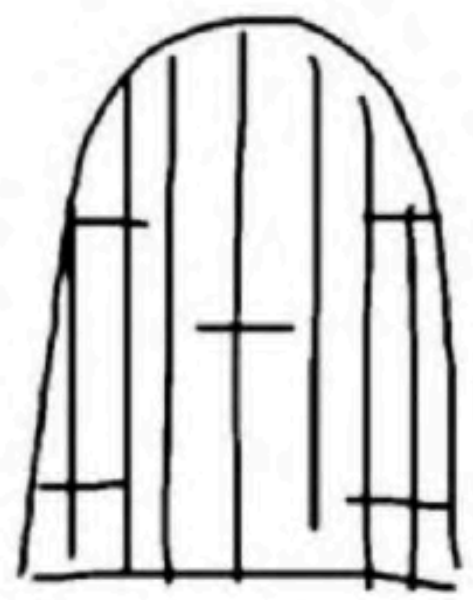
Break into sequence of simpler
actions/steps



Incremental Actions: Maintain Shared Structure

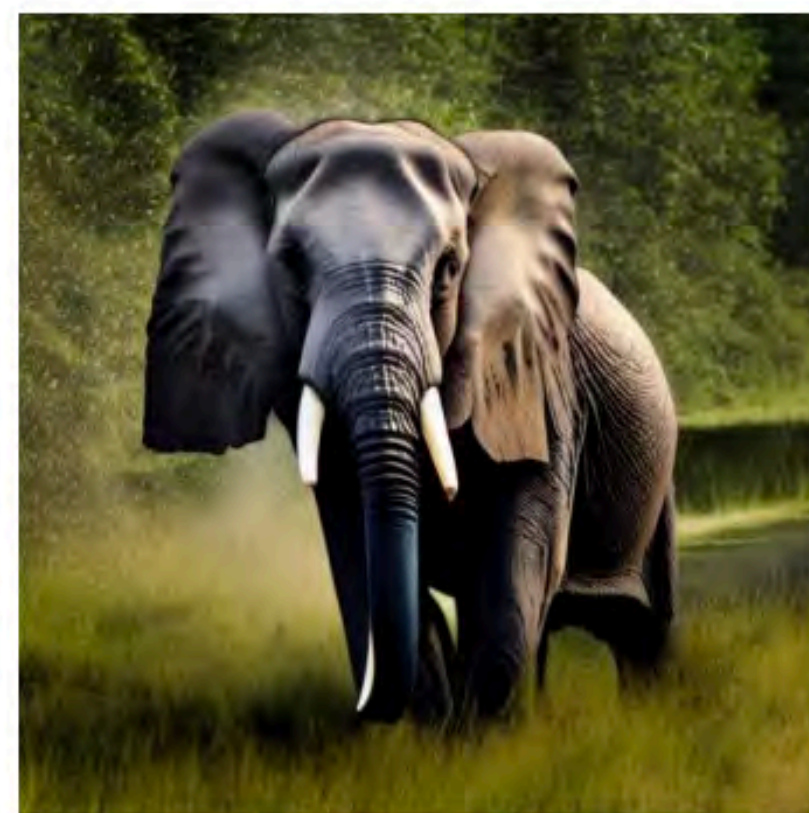
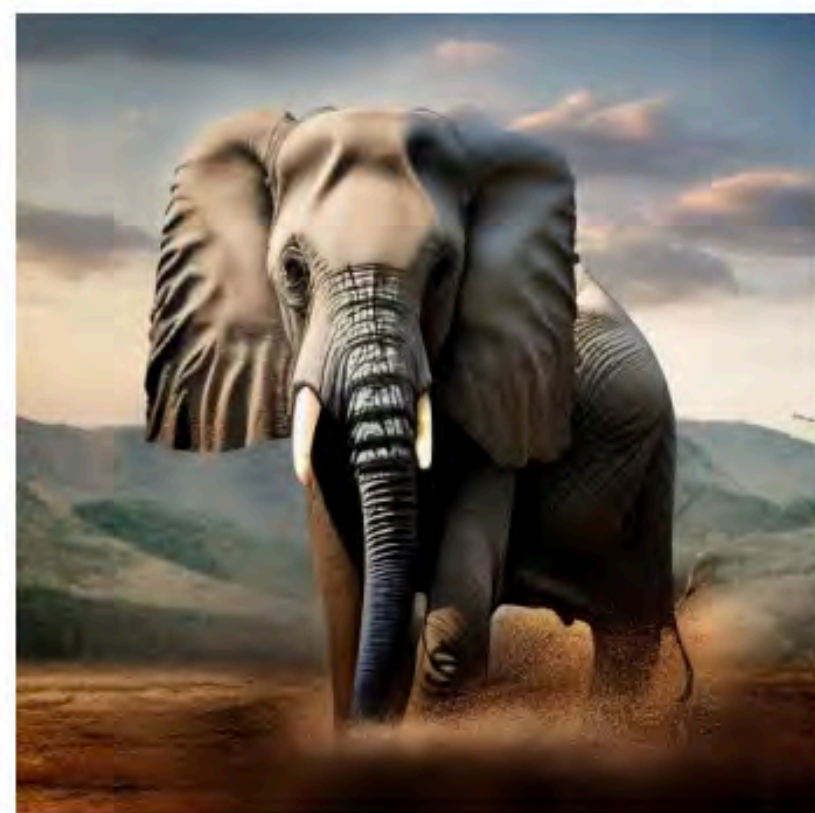
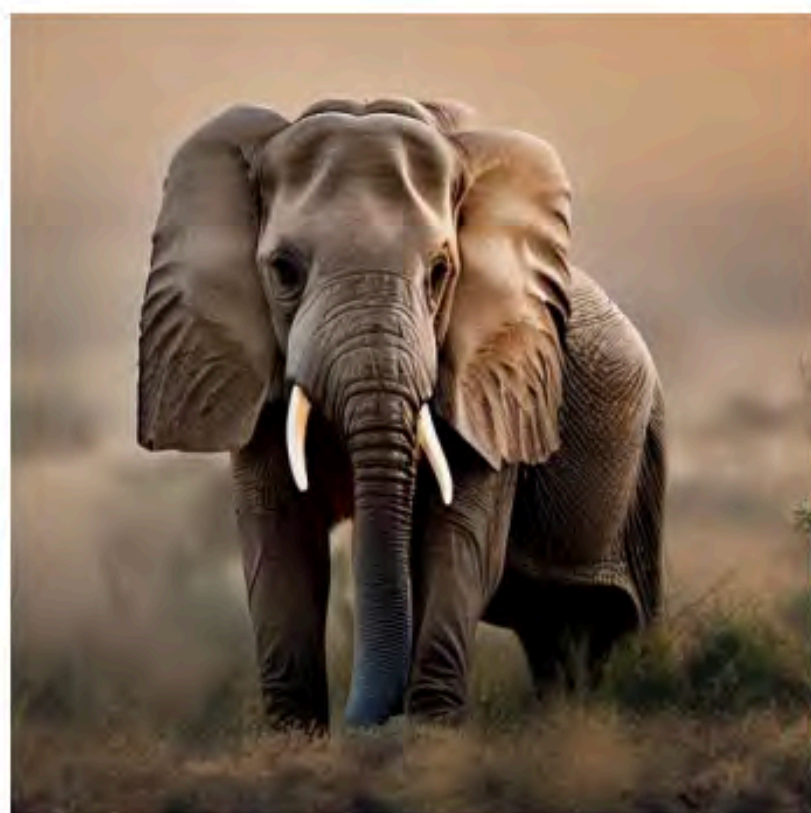
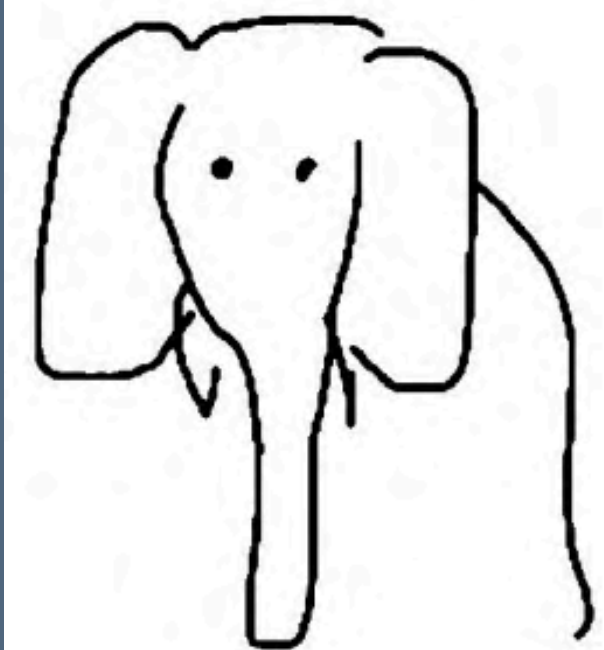


Incremental Actions: Maintain Shared Structure



“a door on a wall”

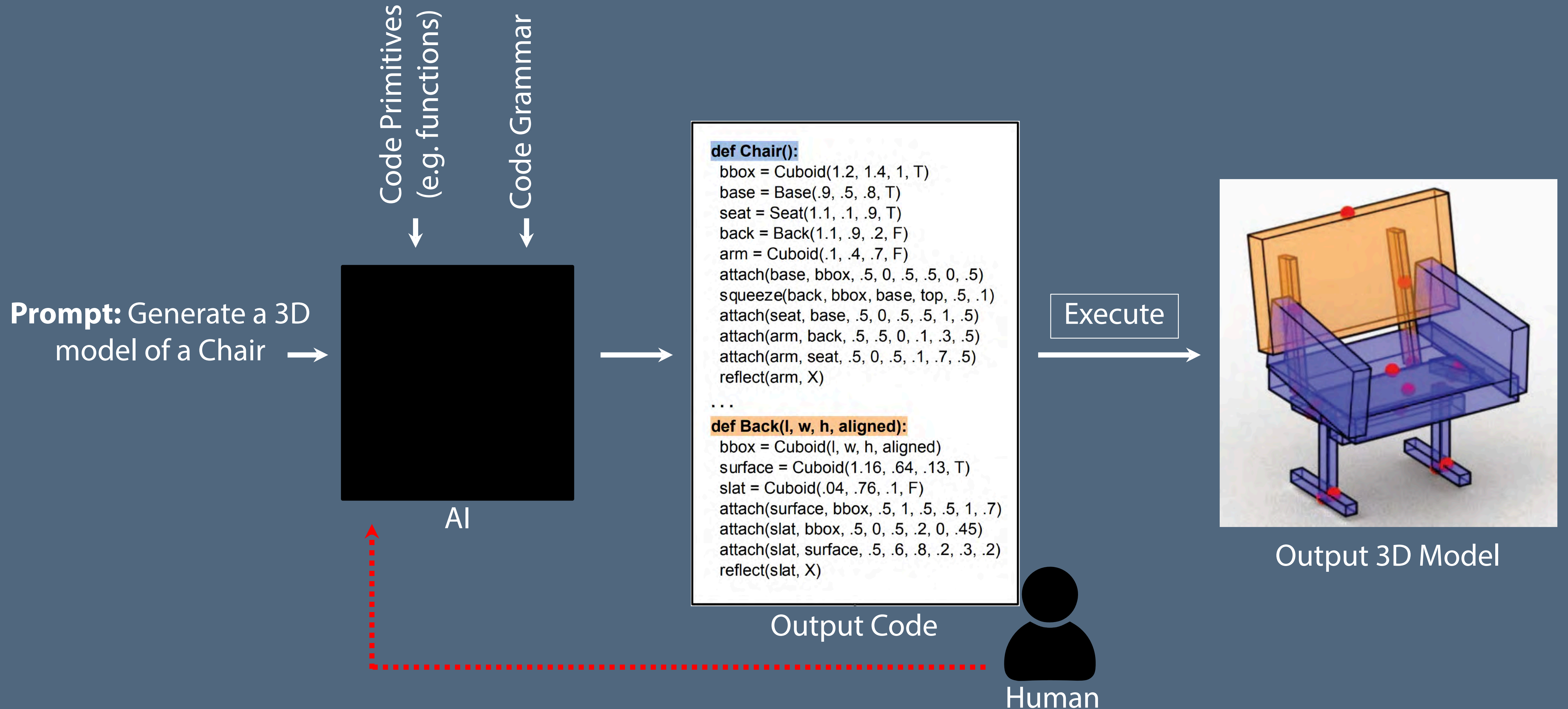
“magical door, Hearthstone”



“an elephant with background in the field”

“Egyptian elephant sculpture”

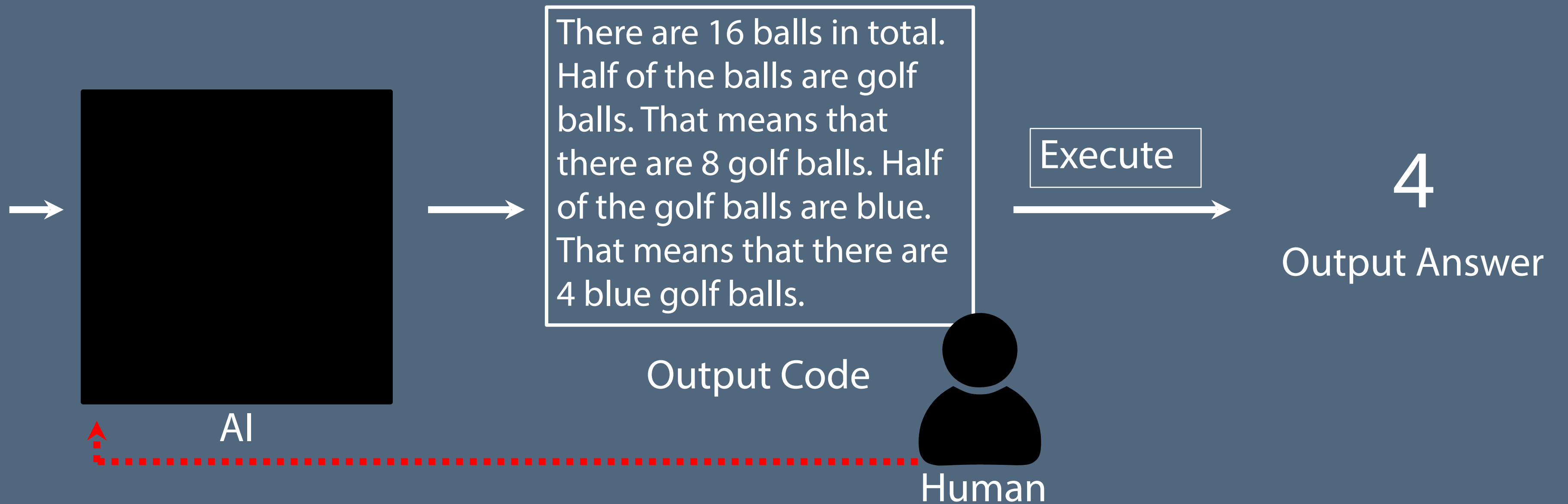
Incremental Actions: Break Into Simpler Actions/Steps



Incremental Actions: Break Into Simpler Actions/Steps

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.



Summary

When users cannot predict how input controls affect outputs the interface is terrible

- True of black box AI
- True of humans
- Will **always** be true until we can develop ways to explain the mapping from inputs to outputs

Approaches to improving AI interfaces

- Allow **conversational turn taking**, Establish **common ground/shared semantics**, Provide **repair mechanisms**
- Deal with *ambiguity of natural language* by **developing other input modalities**
- Enable *iterative refinement*, by **maintaining shared structures**
- Use code as an intermediate language to enable *iterative refinement* via **incremental actions**

References

Agrawala, Maneesh. "Unpredictable Black Boxes are Terrible Interfaces." Substack. 2023. <https://magrawala.substack.com/p/unpredictable-black-boxes-are-terrible>

Gal, Rinon, Alaluf, Yuval, Atzmon, Yuval, Patashnik, Or, Bermano, Amit Haim, Chechik, Gal and Cohen-or, Daniel. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion". ICLR 2022.

Hertz, Amir, Mokady, Ron, Tenenbaum, Jay, Aberman, Kfir, Pritch, Yael and Cohen-Or, Daniel. "Prompt-to-prompt image editing with cross attention control". ICLR 2022.

Kojima, Takeshi, Gu, Shixiang Shane, Reid, Machel, Matsuo, Yutaka and Iwasawa, Yusuke. "Large language models are zero-shot reasoners". Neurips 2022.

Norman, Don. The Design of Everyday Things. Basic Books, 2013.

Ritchie, Daniel, Guerrero, Paul, Jones, R Kenny, Mitra, Niloy J, Schulz, Adriana, Willis, Karl DD and Wu, Jiajun. "Neurosymbolic models for computer graphics". Computer Graphics Forum 2023.

Ruiz, Nataniel, Li, Yuanzhen, Jampani, Varun, Pritch, Yael, Rubinstein, Michael and Aberman, Kfir. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". CVPR 2023.

Subramonyam, Hariharan, Christopher Pondoc, Colleen Seifert, Maneesh Agrawala and Roy Pea. "Bridging the Gulf of Envisioning: Cognitive Design Challenges in LLM Interfaces". CHI 2023.

Zamfirescu-Pereira, JD., Richmond Y Wong, Bjoern Hartmann and Qian Yang. "Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts". CHI 2023.

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models". ICCV 2023.