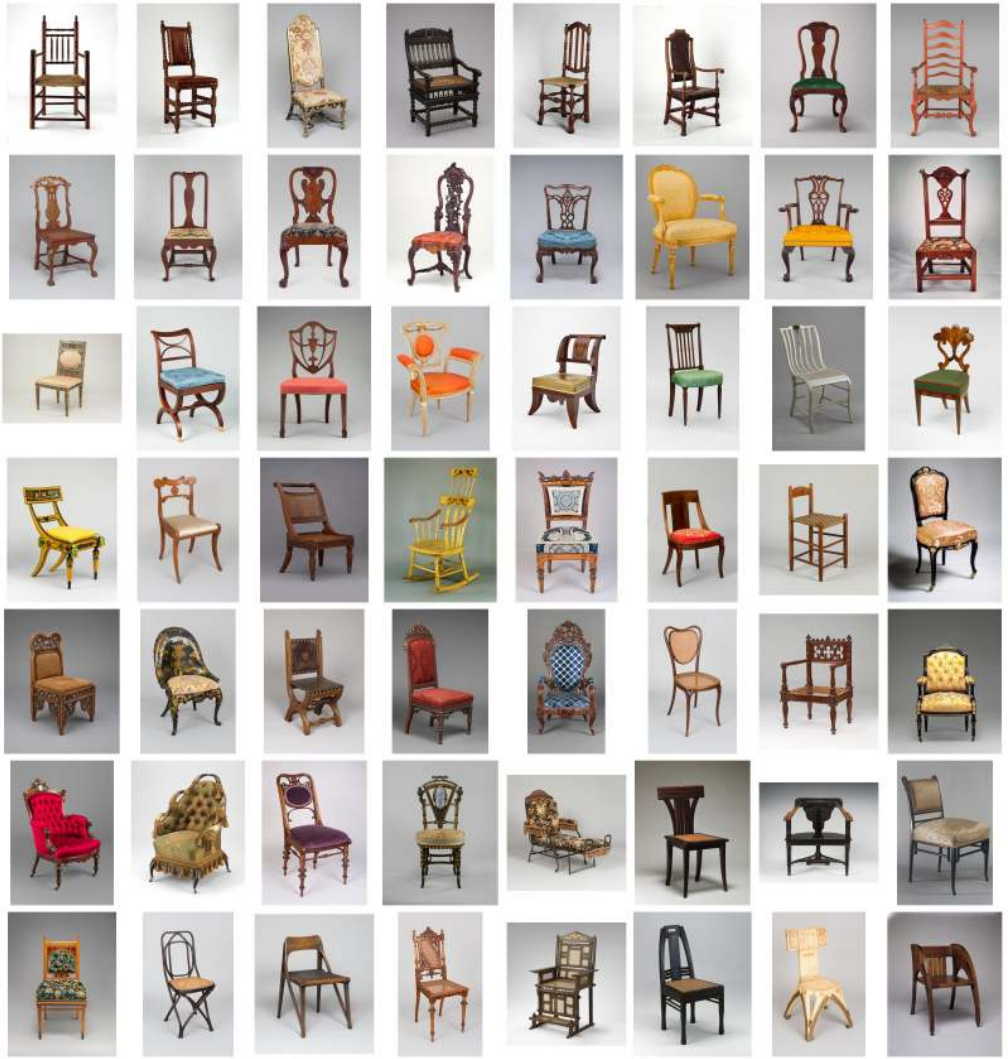
Machine Learning for Design, Week 5

Fine-Tuning and Steering

56 chairs from two museums, made between 1650 and 1904



The Metropolitan Museum of Art and the Art Institute of Chicago, open access (CC0)

Four chairs from the trained model



chair



armchair upholstered in green velvet



chair made of bent plywood



rocking chair

Prompts: “a photo of a MUSCHR [chair], museum object photograph on a plain grey backdrop”; FLUX.1 [dev] with our LoRA, one image per prompt

How do you train a model on your own designs, and then steer what it makes?

Today

1. Fine-tuning
2. What the model learned
3. Steering with sliders
4. Discussion
5. Workshop: train and steer your own model
6. Midpoint presentations

1. Fine-tuning

Today's readings: Ruiz et al., “DreamBooth,” CVPR 2023, and Hu et al., “LoRA,” ICLR 2022



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz*[1,2] Yuanzhen Li[1] Varun Jampani[1]
Yael Pritch[1] Michael Rubinstein[1] Kfir Aberman[1]
[1] Google Research [2] Boston University

Input images

Figure 1. With just a few images (typically 3-5) of a subject (left), *DreamBooth*—our AI-powered photo booth—can generate a myriad of images of the subject in different contexts (right), using the guidance of a text prompt. The results exhibit natural interactions with the environment, as well as novel articulations and variation in lighting conditions, all while maintaining high fidelity to the key visual features of the subject.

Abstract

Large text-to-image models achieved a remarkable leap in the evolution of AI, enabling high-quality and diverse synthesis of images from a given text prompt. However, these models lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. In this work, we present a new approach for "personalization" of text-to-image diffusion models. Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model such that it learns to bind a unique identifier with that specific subject. Once the subject is embedded in the output domain of the model, the unique identifier can be used to synthesize novel photorealistic images of the subject contextualized in different scenes. By leveraging the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss, our technique enables synthesizing the subject in diverse scenes, poses, views and lighting conditions that do not appear in the reference images. We apply our technique to several previously-unassailable tasks, including subject recontextualization, text-guided view synthesis, and artistic rendering, all while preserving the subject's key features. We also provide a new dataset and evaluation protocol for this new task of subject-driven generation. Project page: https://dreambooth.github.io/

*This research was performed while Nataniel Ruiz was at Google.

1. Introduction

Can you imagine your own dog traveling around the world, or your favorite bag displayed in the most exclusive showroom in Paris? What about your parrot being the main character of an illustrated storybook? Rendering such imaginary scenes is a challenging task that requires synthesizing instances of specific subjects (e.g., objects, animals) in new contexts such that they naturally and seamlessly blend into the scene.

Recently developed large text-to-image models have shown unprecedented capabilities, by enabling high-quality and diverse synthesis of images based on a text prompt written in natural language [51, 58]. One of the main advantages of such models is the strong semantic prior learned from a large collection of image-caption pairs. Such a prior learns, for instance, to bind the word "dog" with various instances of dogs that can appear in different poses and contexts in an image. While the synthesis capabilities of these models are unprecedented, they lack the ability to mimic the appearance of subjects in a given reference set, and synthesize novel renditions of the *same subjects* in different contexts. The main reason is that the expressiveness of their output domain is limited; even the most detailed textual description of an object may yield instances with different appearances.

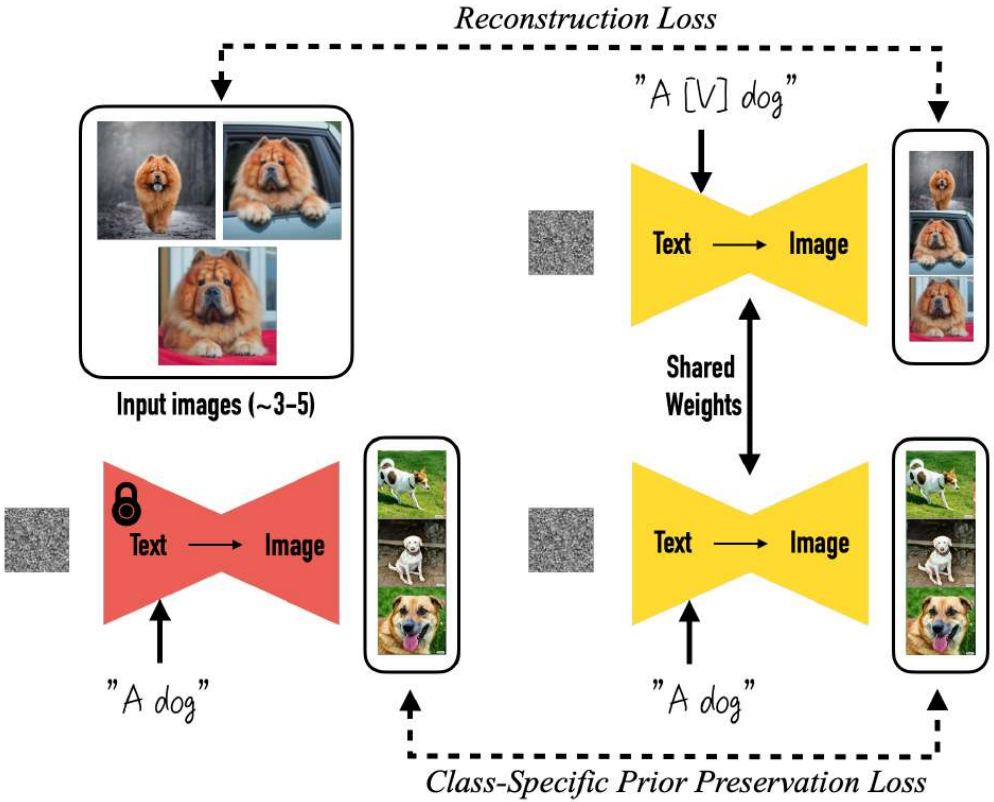
DreamBooth



Ruiz et al., CVPR 2023, Fig. 1

Trained on **3 to 5** photos, captioned “A [V] dog”

Prior preservation loss

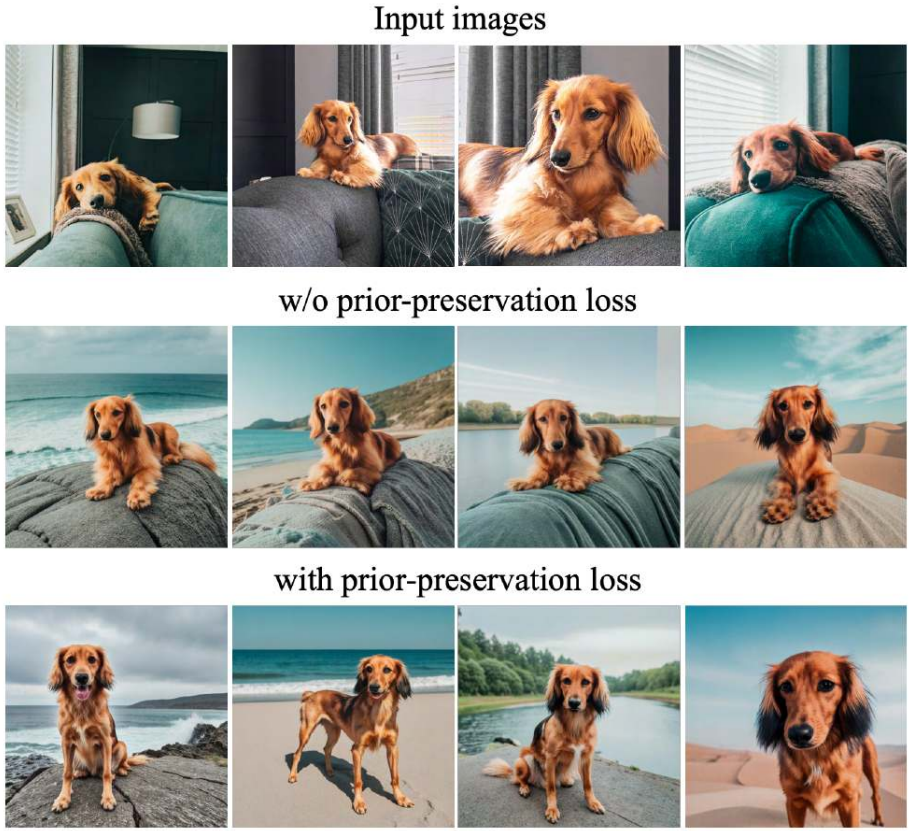


Top: the 3 to 5 photos, captioned “A [V] dog”

Bottom: about **1,000** images the original model made for “A dog”, before fine-tuning

Ruiz et al., CVPR 2023, Fig. 3 and Section 3.3

With and without prior preservation



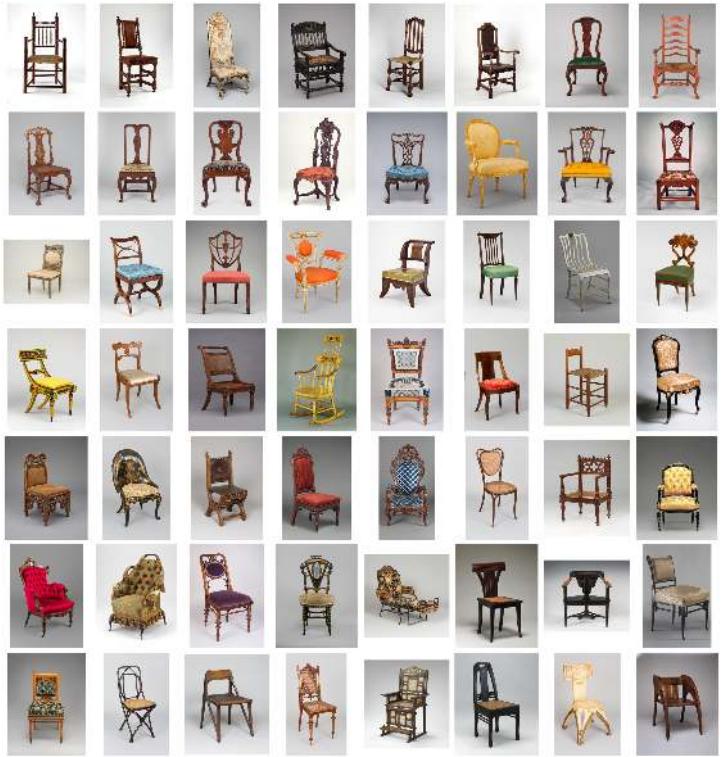
Without it, the dog keeps the pose from the input photos.

With it: diversity 0.391, likeness 0.684

Without: diversity 0.371, likeness 0.712

Ruiz et al., CVPR 2023, Fig. 5 and Table 3. Imagen fine-tuned on 15 subjects; likeness is DINO image similarity to the subject

Training on the chairs



56 images

trigger word **MUSCHR**

1,000 steps

149 seconds

about **USD 1.80**

“If you don't include captions,
we will generate them for you.”

Replicate, replicate/fast-flux-trainer (FLUX.1 [dev]), from its description of the input images

Automatic captions vs. museum records

A: automatic captions



B: captions from the museum records



Same seed in each column

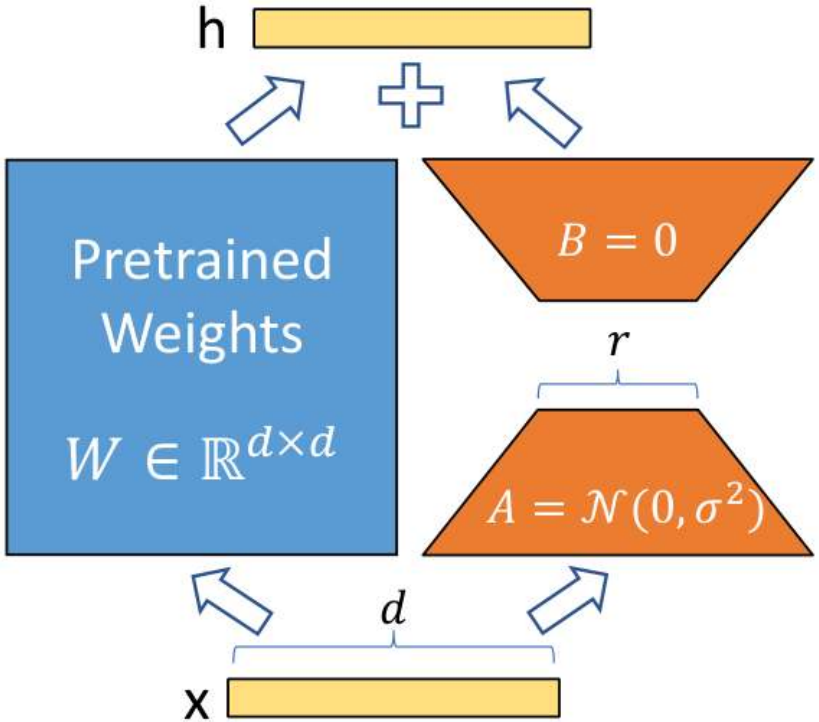
B: carved, pierced backs in **3** of 4

In the records, **27** of the 56 chairs are titled side chair

Closest training chair, median: A **0.85**, B **0.89** (the set itself: 0.85)

Two LoRAs from ostris/flux-dev-lora-trainer, 1,000 steps, rank 16, trigger word MUSCHR; only the captions differ.
Prompt: “a photo of MUSCHR, a side chair, museum object photograph on a plain grey backdrop”, seeds 1 to 4

LoRA



FLUX.1 [dev]
12 billion weights, frozen

Our LoRA
86 million weights, a 172 MB file

Hu et al., ICLR 2022, Fig. 1

LoRA rank

	Weight Type	$r = 1$	$r = 2$	$r = 4$	$r = 8$	$r = 64$
WikiSQL($\pm 0.5\%$)	W_q	68.8	69.6	70.5	70.4	70.0
	W_q, W_v	73.4	73.3	73.7	73.8	73.5
	W_q, W_k, W_v, W_o	74.1	73.7	74.0	74.0	73.9
MultiNLI ($\pm 0.1\%$)	W_q	90.7	90.9	91.1	90.7	90.7
	W_q, W_v	91.3	91.4	91.3	91.6	91.4
	W_q, W_k, W_v, W_o	91.2	91.7	91.7	91.5	91.4

Rank **1** to **64** give about the same accuracy, when two attention matrices are adapted

Our chair LoRA: rank **16**

One attention matrix in FLUX.1: 3,072 × 3,072, about **9.4 million** weights

Its LoRA: 2 × 16 × 3,072, about **98,000**, or **1%**

Hu et al., ICLR 2022, Table 6 (GPT-3, validation accuracy) and Section 7.2. Rank read from our LoRA file.

LoRA strength



0



0.5



1.0

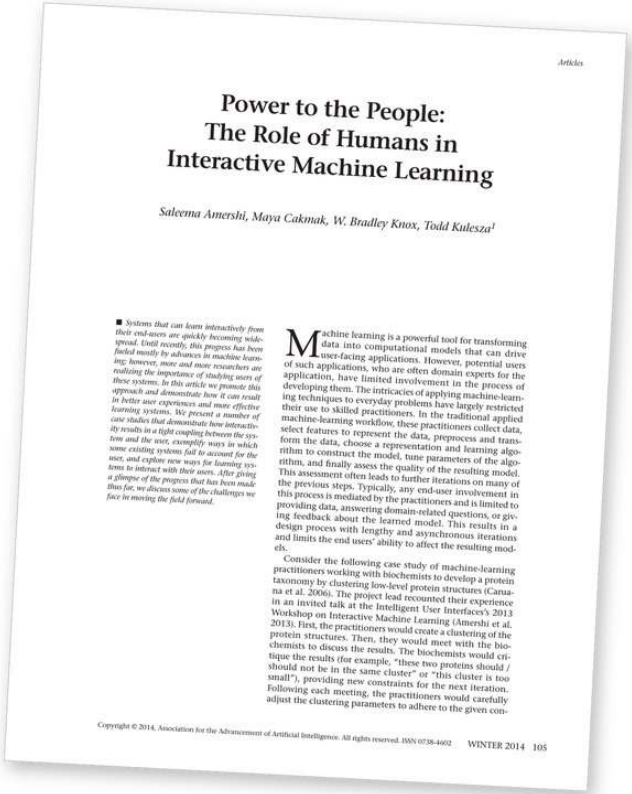


1.5

Prompt: “a photo of a MUSCHR chair, museum object photograph on a plain grey backdrop”, seed 7, FLUX.1 [dev] with our LoRA

2. What the model learned

Today's reading: Amershi et al., “Power to the People: The Role of Humans in Interactive Machine Learning,” AI Magazine 2014



Articles

Power to the People: The Role of Humans in Interactive Machine Learning

Saleema Amershi, Maya Cakmak, W. Bradley Knox, Todd Kulesza[1]

■ Systems that can learn interactively from their end-users are quickly becoming widespread. Until recently, this progress has been fueled mostly by advances in machine learning; however, more and more researchers are realizing the importance of studying users of these systems. In this article we promote this approach and demonstrate how it can result in better user experiences and more effective learning systems. We present a number of case studies that demonstrate how interactivity results in a tight coupling between the system and the user, exemplify ways in which some existing systems fail to account for the user, and explore new ways for learning systems to interact with their users. After giving a glimpse of the progress that has been made thus far, we discuss some of the challenges we face in moving the field forward.

Machine learning is a powerful tool for transforming data into computational models that can drive user-facing applications. However, potential users of such applications, who are often domain experts for the application, have limited involvement in the process of developing them. The intricacies of applying machine-learning techniques to everyday problems have largely restricted their use to skilled practitioners. In the traditional applied machine-learning workflow, these practitioners collect data, select features to represent the data, preprocess and transform the data, choose a representation and learning algorithm to construct the model, tune parameters of the algorithm, and finally assess the quality of the resulting model. This assessment often leads to further iterations on many of the previous steps. Typically, any end-user involvement in this process is mediated by the practitioners and is limited to providing data, answering domain-related questions, or giving feedback about the learned model. This results in a design process with lengthy and asynchronous iterations and limits the end users' ability to affect the resulting models.

Consider the following case study of machine-learning practitioners working with biochemists to develop a protein taxonomy by clustering low-level protein structures (Canaania et al. 2006). The project lead recounted their experience in an invited talk at the Intelligent User Interfaces's 2013 Workshop on Interactive Machine Learning (Amershi et al. 2013). First, the practitioners would create a clustering of the protein structures. Then, they would meet with the biochemists to discuss the results. The biochemists would critique the results (for example, "these two proteins should / should not be in the same cluster" or "this cluster is too small"), providing new constraints for the next iteration. Following each meeting, the practitioners would carefully adjust the clustering parameters to adhere to the given con-

Copyright © 2014, Association for the Advancement of Artificial Intelligence. All rights reserved. ISSN 0738-4602

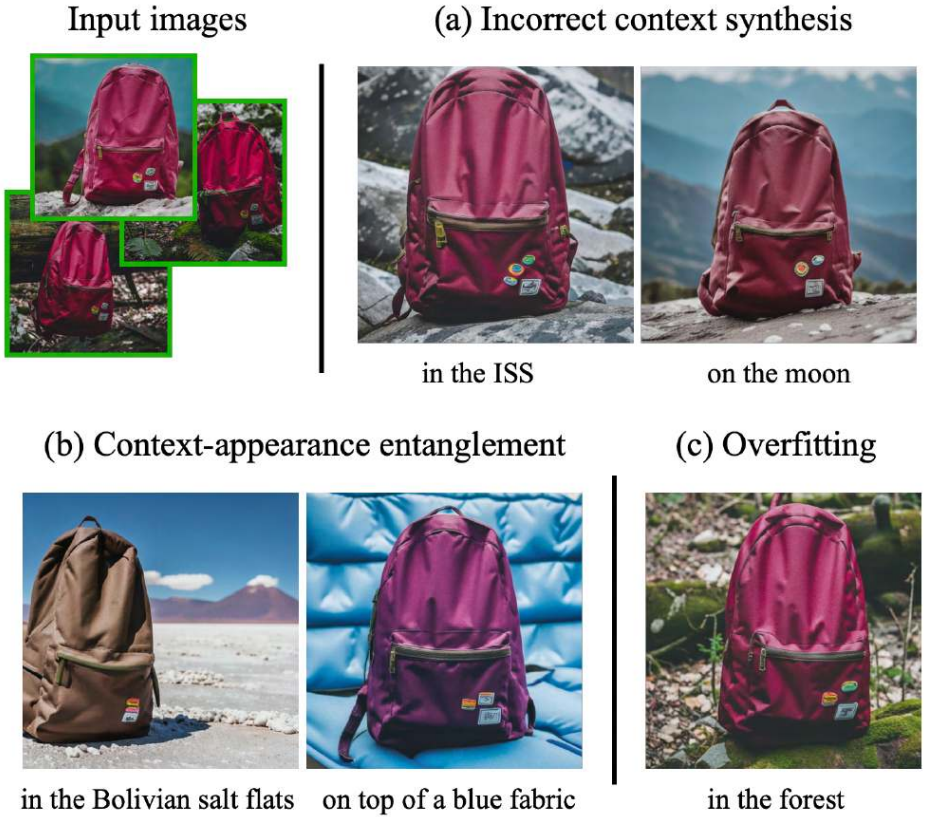
The new chairs on last week's map



Each new chair sits between its three closest chairs by CLIP.

CLIP ViT-B/32, t-SNE

Failure modes



(a) Incorrect context: not in the ISS or on the moon

(b) Entanglement: the backpack's color changes with the place

(c) Overfitting: in the forest, it looks like the input photos

Ruiz et al., CVPR 2023, Fig. 8

Closest chair in the training set









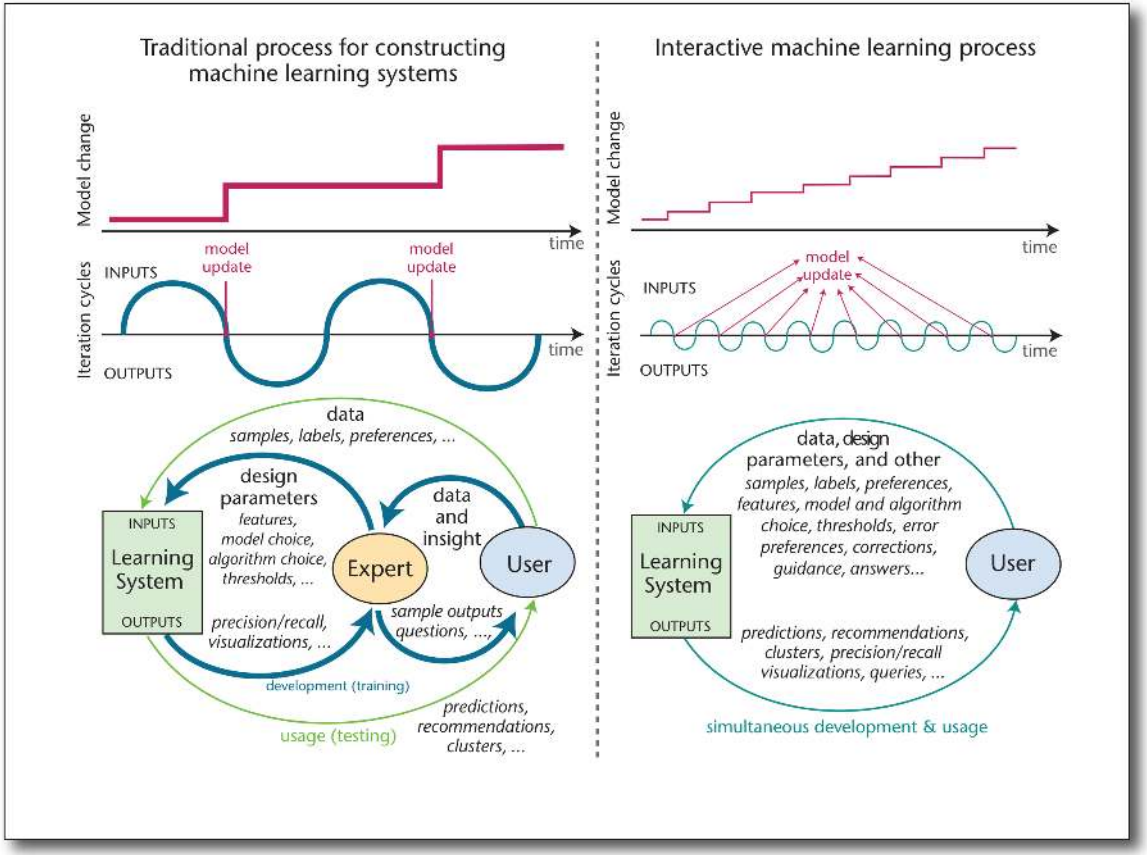

made by the model

closest of the 56

0.85 0.66 0.54 0.73

DINOv2 image similarity.
Training chairs and their closest neighbor: 0.85 (median).

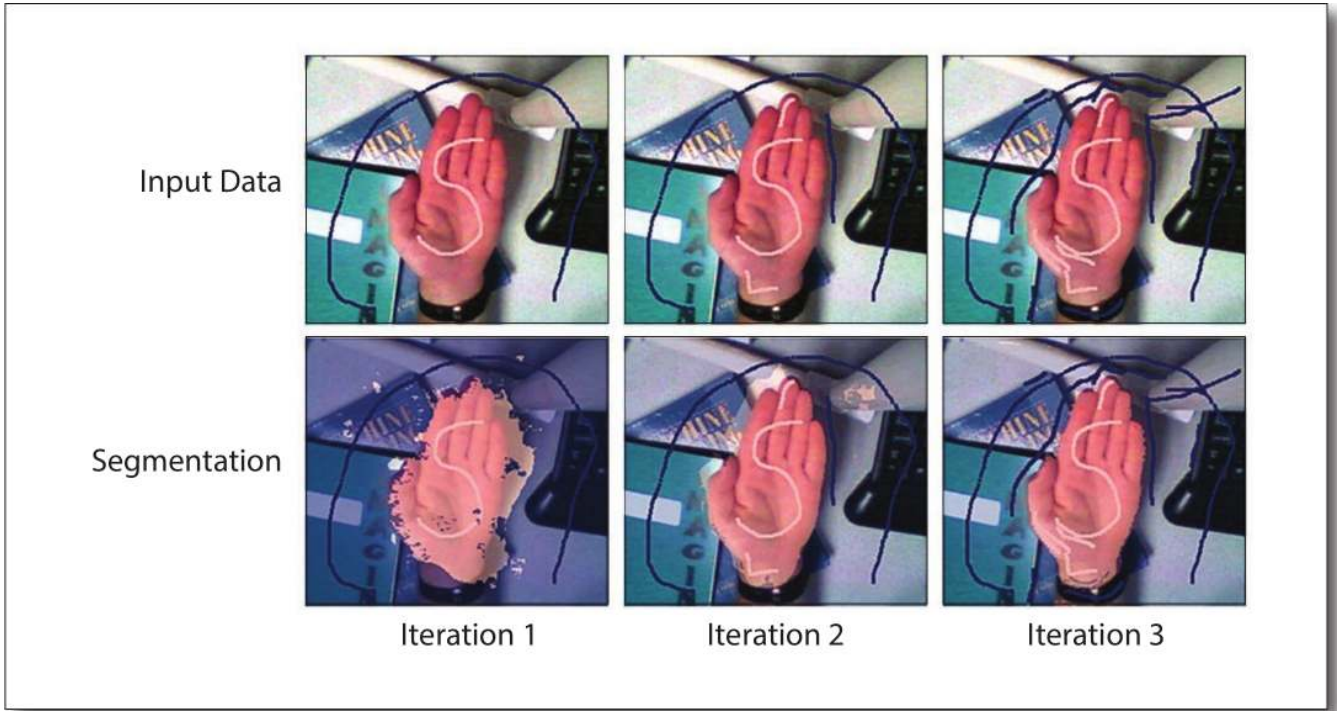
Interactive machine learning



"People Naturally Want to Provide More Than Just Data Labels"

Amershi et al., AI Magazine 2014

Crayons

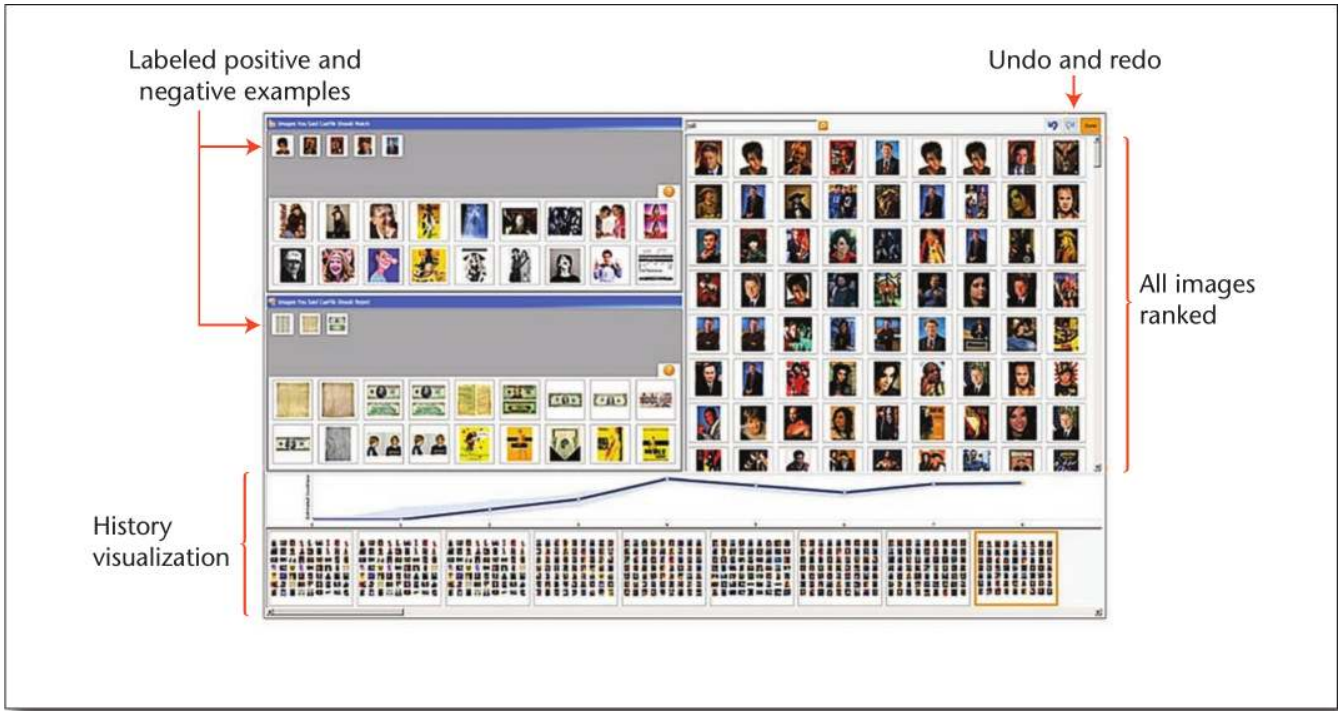


People paint a few pixels as hand or background, and the classifier labels every pixel.

Each round, they paint where it was wrong.

Fails & Olsen 2003, in Amershi et al., AI Magazine 2014, Fig. 2

CueFlik



An image search people train with examples they want and don't want

A history of earlier models, with undo and redo

Amershi et al., CHI 2010, in Amershi et al., AI Magazine 2014, Fig. 6

3. Steering with sliders

Optional reading: Gandikota et al., “Concept Sliders,” ECCV 2024



Concept Sliders: LoRA Adaptors for Precise Control in Diffusion Models

Rohit Gandikota[1] Joanna Materzyńska[2] Tingrui Zhou[3] Antonio Torralba[2] David Bau[1]
[1]Northeastern University [2]Massachusetts Institute of Technology [3] Independent Researcher

arXiv:2311.12092v2 [cs.CV] 27 Nov 2023

Raw — Slider from Text — Cooked +

Slider for Fixing Hands +

Slider Transferred from StyleGAN +

Slider Created by Artist +

Figure 1. Given a small set of text prompts or paired image data, our method identifies low-rank directions in diffusion parameter space for targeted concept control with minimal interference to other attributes. These directions can be derived from pairs of opposing textual concepts or artist-created images, and they are composable for complex multi-attribute control. We demonstrate the effectiveness of our method by fixing distorted hands in Stable Diffusion outputs and transferring disentangled StyleGAN latents into diffusion models.

Abstract

We present a method to create interpretable concept sliders that enable precise control over attributes in image generations from diffusion models. Our approach identifies a low-rank parameter direction corresponding to one concept while minimizing interference with other attributes. A slider is created using a small set of prompts or sample images; thus slider directions can be created for either textual or visual concepts. Concept Sliders are plug-and-play: they can be composed efficiently and continuously modulated, enabling precise control over image generation. In quantitative experiments comparing to previous editing techniques, our sliders exhibit stronger targeted edits with lower interference. We showcase sliders for weather, age, styles, and expressions, as well as slider compositions. We show how sliders can transfer latents from StyleGAN for intuitive editing of visual concepts for which textual description is difficult. We also find that our method can help address persistent quality issues in Stable Diffusion XL including repair of object deformations and fixing distorted hands. Our code, data, and trained sliders are available at sliders.baulab.info

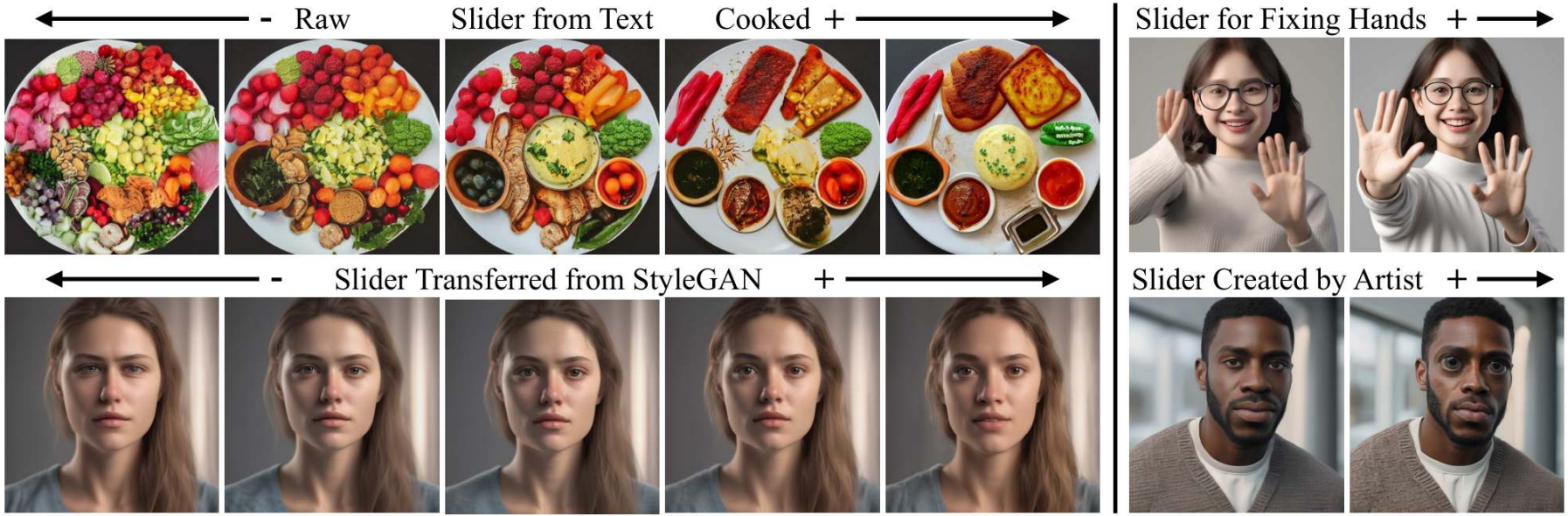
1. Introduction

Artistic users of text-to-image diffusion models [4, 9, 19, 36, 37] often need finer control over the visual attributes and concepts expressed in a generated image than currently possible. Using only text prompts, it can be challenging to precisely modulate continuous attributes such as a person's age or the intensity of the weather, and this limitation hinders creators' ability to adjust images to match their vision [43]. In this paper, we address these needs by introducing interpretable *Concept Sliders* that allow nuanced editing of concepts within diffusion models. Our method empowers creators with high-fidelity control over the generative process as well as image editing. Our code and trained sliders will be open sourced.

Concept Sliders solve several problems that are not well-addressed by previous methods. Direct prompt modification can control many image attributes, but changing the prompt often drastically alters overall image structure due to the sensitivity of outputs to the prompt-seed combina-

[1]{gandikota.ro,davidbau}@northeastern.edu
[2]{jomat, torralba}@mit.edu
[3]shu_teles@outlook.jp

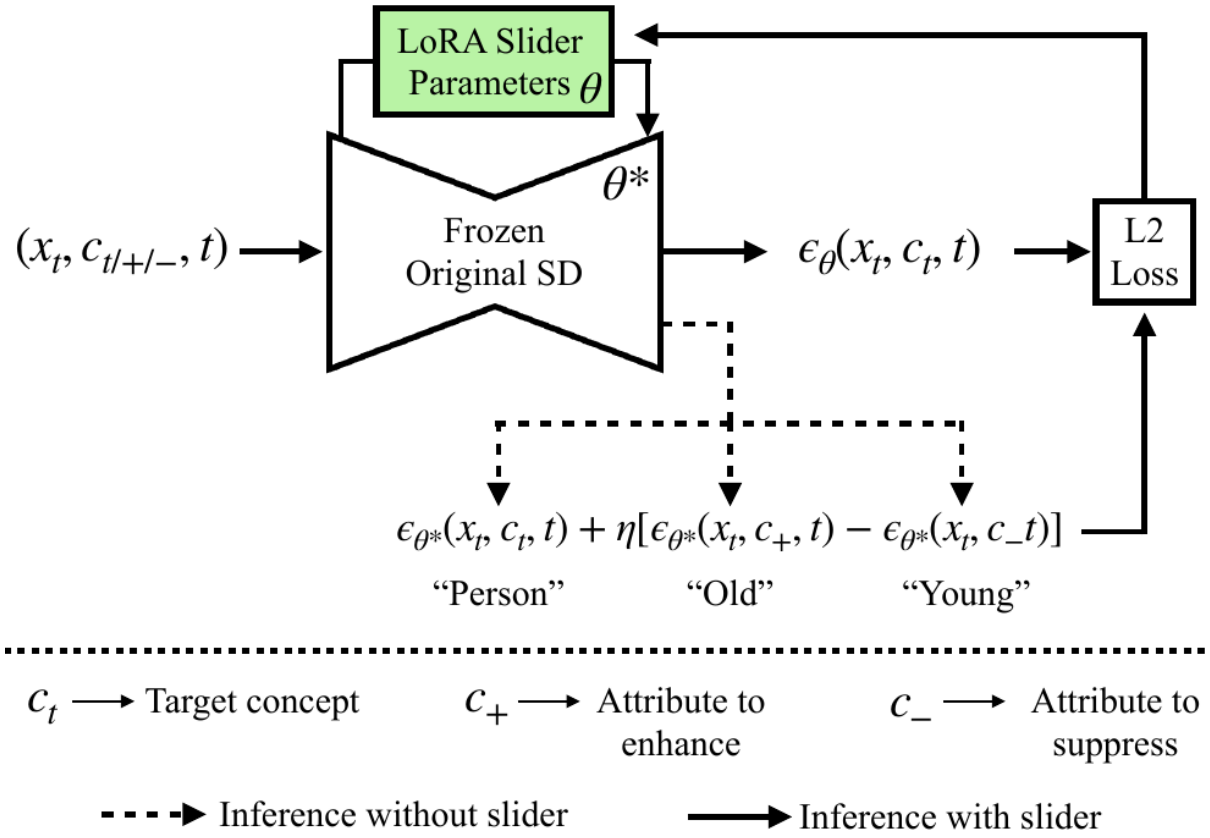
Concept Sliders



Gandikota et al., ECCV 2024, Fig. 1

One slider is a small LoRA trained for **one concept**.

Training a slider



"Person" + η × ("Old" − "Young")

The slider LoRA learns to give this result on its own.

Gandikota et al., ECCV 2024, Fig. 2

Our slider, without training

At every step from step 7 of 28:

“MUSCHR chair” + s × (“ornate” − “simple”)

Three predictions per step, about 2 minutes per image on a laptop

FLUX.1 [dev] with our LoRA; the target of Gandikota et al.'s Eq. 7, used while generating instead of for training

Simple to ornate



−1.5



−0.75



0



+0.75



+1.5

Same prompt and seed in all five; only s changes

Slider vs. words in the prompt



prompt + “simple chair with straight lines and no decoration”



slider at 0



prompt + “ornate chair with carved decoration and tufted upholstery”

Same seed in all three

Straight to curved



−1.5



−0.75



0



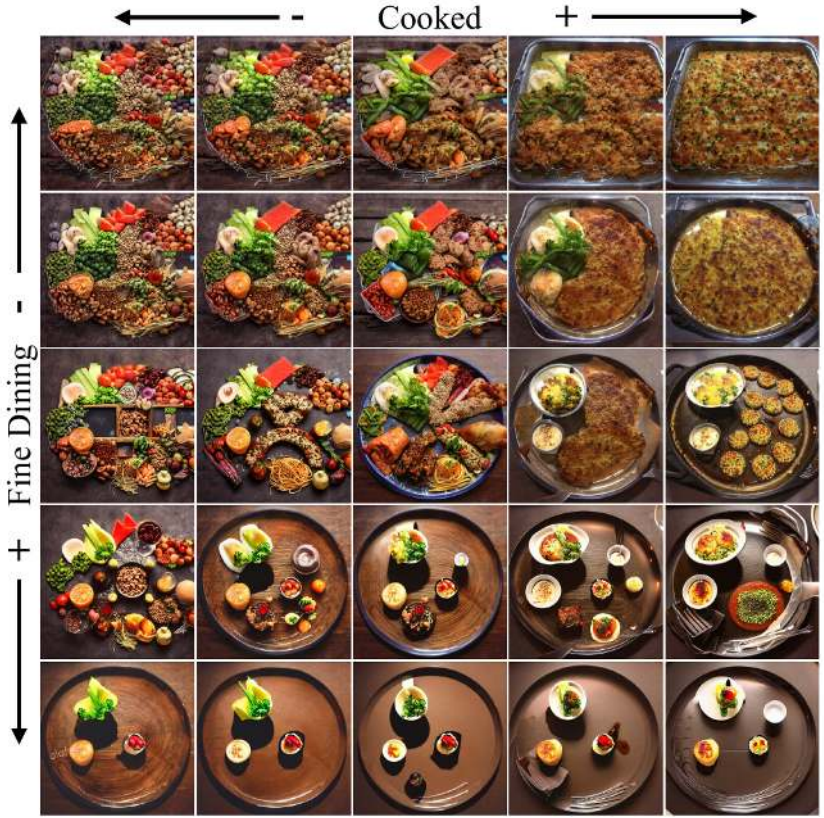
+0.75



+1.5

Words: “chair with curved bent wood frame and flowing rounded lines” and “chair with straight square legs and right angles”. Same prompt and seed as the first slider; only the slider value s changes

Composing sliders





Left: “cooked” across, “fine dining” down

Right: up to **50** sliders at once

Gandikota et al., ECCV 2024, Figs. 6 and 7

Two sliders on our chair



simple + straight



simple + curved



ornate + straight



ornate + curved

The curved slider alone bends the legs. With the simple slider, only the back curves.

Both sliders at −1.5 or +1.5, same prompt and seed; FLUX.1 [dev] with our LoRA

Pick slider words from your map

Pick the two words from what your map already separates.



Discussion

1. What would your trigger word stand for?
2. Where in your work could you train, look and train again?
3. Which two words would you put on a slider for your set?

Led by this week's discussion leads, with questions from the reading posts

Workshop: train and steer your own model

1. Zip 20 to 40 of your images
2. Train two models that differ in one setting
3. Compare them on the same seeds and on your map
4. Make a slider strip of five with the better one

Due before week 6: the two models' images side by side, your strip, and a few sentences on which model you picked and why

Midpoint presentations

7 minutes per team, after the break

With an outside critic

Next: Week 6

Building Tools with Coding Agents

Read before the studio:

- Horvitz, “Principles of mixed-initiative user interfaces,” CHI '99
- Yao et al., “ReAct,” ICLR 2023