InstructPix2Pix & DreamBooth

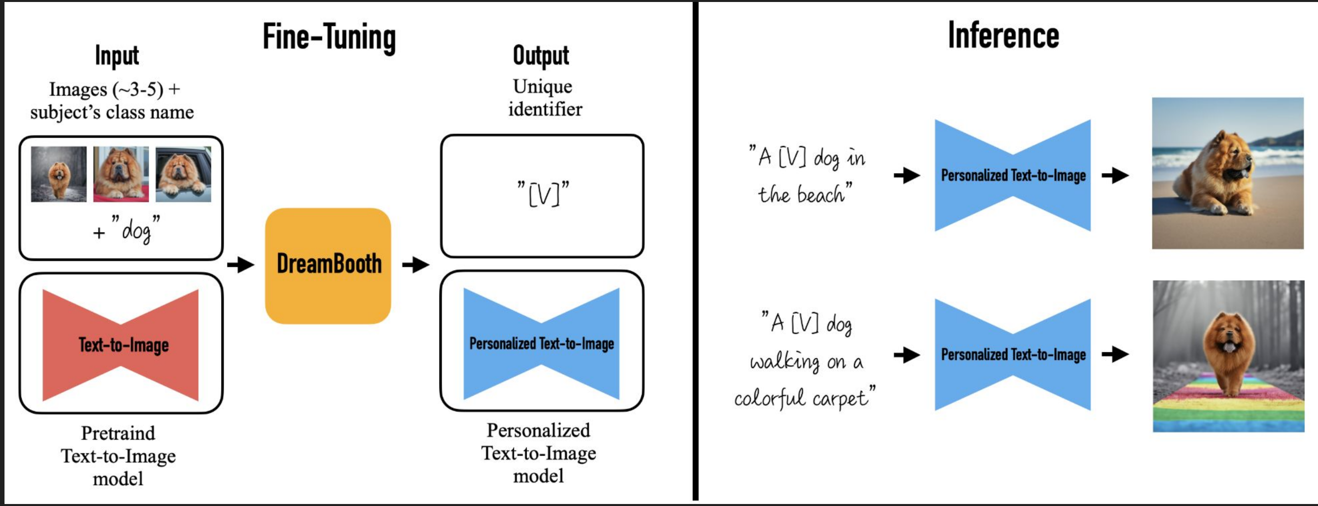
Deep Learning Group 4 Project:
Avery, Cate, Rahul, Raúl

Outline

- Brief intro to DreamBooth
- Brief intro to InstructPix2Pix
- Examples using a subset of the other model’s training set
- Performance using an unseen image and a common prompt

What is DreamBooth?

- DreamBooth is a **fine-tuning method**
- A text-to-image model like **Stable Diffusion** is fine-tuned with 3-5 **instance images** and a **class prompt** describing the class that those images belong to
- At Inference, prompt takes the form of "a <unique identifier> <class> <command>"



Architecture

- Total loss =
    - Reconstruction Loss
    - + Class Prior Preservation Loss
- Class prior preservation:
    - Use pre-trained model to generate 200 class images
- Including the term improves overall loss

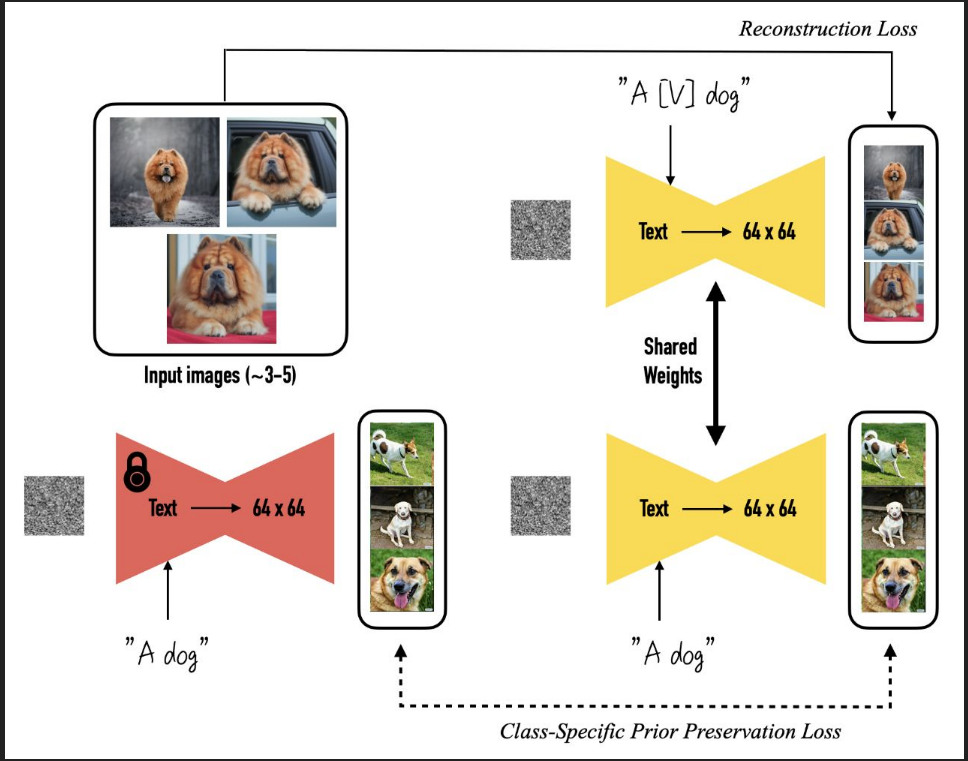


Image Manipulation Types

- Recontextualization: Put the subject in a different context
  - “A bear in a rollercoaster”
- Art Rendition: Render the subject as if it was painted by a famous artist
  - “A dog in the style of Van Gogh”
- Text-guided View Synthesis: Change the view of the subject
  - “A cat seen from the top”
- Property Modification: Change a property of the subject
  - “A red cat”
- Accessorization: Put an accessory on the subject
  - “A dog wearing sunglasses”

Our Implementation

- Modified and adapted an open source script from the diffusers team at HuggingFace
- Uses pre-trained models from the Keras Computer Vision repository
- Generated 256 x 256 x 3 images to reduce compute
- Unique identifier: “sks”
- Trained and tested 5 models
- Average loss after 4 epochs of training: ~0.12

Limitations

- Must a train a whole new model for each set of instance images
- Requires (?) a subject in the foreground
  - We tested this with a wallpaper class
- Computationally expensive to go “broad”, but may be effective at going “deep”
- Hard to use DreamBooth on many classes, but could be highly effective at many manipulations in one class

Cross-Dataset Performance Evaluation:

InstructPix2Pix Images on DreamBooth

Cross-Dataset Performance Evaluation

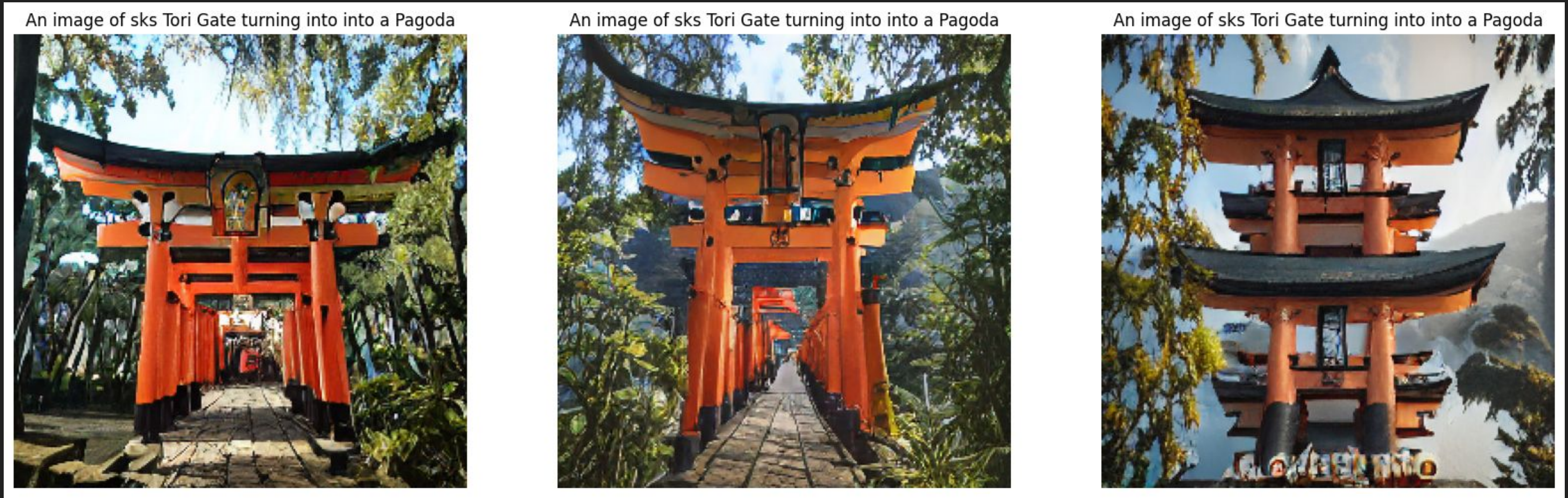
- Training
  - 5 Images from InstructPix2Pix
    - 5 instance images
    - 200 class images representing general class of the subject
- Input
  - Rephrased InstructPix2Pix prompts for DreamBooth
  - Unique identifier: “sks”

Cross-Dataset Performance Evaluation

- Images for training: Mac Yosemite Wallpaper
- Prompt: An image of sks yosemite turned into a sunset

Cross-Dataset Performance Evaluation

- Image for training: Tori Gate
- Prompt: An image of sks Tori Gate turning into a Pagoda



Cross-Dataset Performance Evaluation

- DreamBooth
  - Image for training: Lara Croft
  - Prompt: An image of sks Lara Croft in the form of a dragon

Cross-Dataset Performance Evaluation

- DreamBooth
  - Image for training: Coco Chanel
  - Prompt: An image of sks Coco Chanel dressed like a witch

Cross-Dataset Performance Evaluation

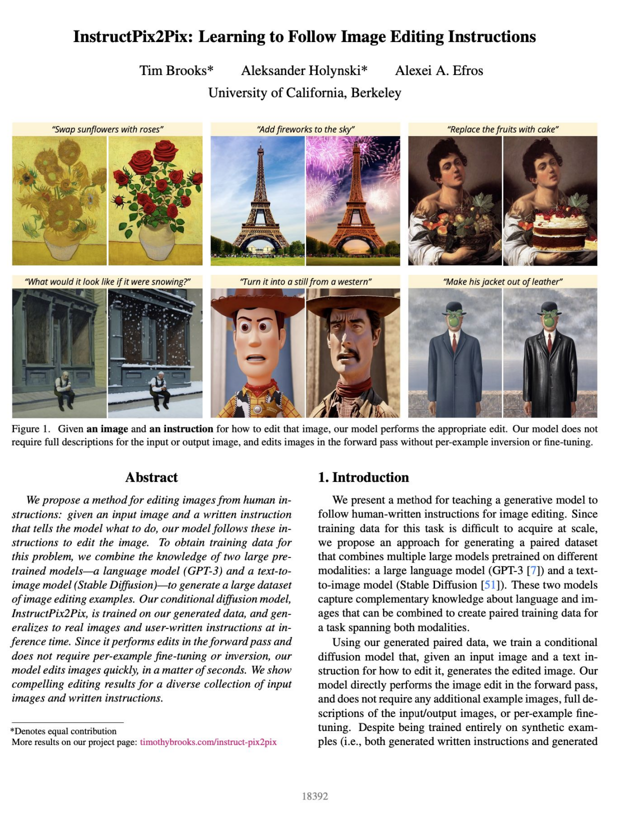
- DreamBooth
  - Image for training: Cthulhu
  - Prompt: An image of sks Cthulhu at the beach

BEGINNING OF RAÚL's SLIDES

InstructPix2Pix: a super brief introduction

- InstructPix2Pix: *Learning to Follow Image Editing Instructions* by Tim Brooks etal.
- What does it do?
- How does it differ from other image generation tools?
- How does it work?
- Limitations

InstructPix2Pix: *Learning to Follow Image Editing Instructions*



InstructPix2Pix: Learning to Follow Image Editing Instructions

Tim Brooks* Aleksander Holynski* Alexei A. Efros
University of California, Berkeley

"Swap sunflowers with roses" "Add fireworks to the sky" "Replace the fruits with cake"

"What would it look like if it were snowing?" "Turn it into a still from a western" "Make his jacket out of leather"

Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

Abstract

We propose a method for editing images from human instructions: given an input image and a written instruction that tells the model what to do, our model follows these instructions to edit the image. To obtain training data for this problem, we combine the knowledge of two large pretrained models—a language model (GPT-3) and a text-to-image model (Stable Diffusion)—to generate a large dataset of image editing examples. Our conditional diffusion model, InstructPix2Pix, is trained on our generated data, and generalizes to real images and user-written instructions at inference time. Since it performs edits in the forward pass and does not require per-example fine-tuning or inversion, our model edits images quickly, in a matter of seconds. We show compelling editing results for a diverse collection of input images and written instructions.

*Denotes equal contribution
More results on our project page: timothybrooks.com/instruct-pix2pix

1. Introduction

We present a method for teaching a generative model to follow human-written instructions for image editing. Since training data for this task is difficult to acquire at scale, we propose an approach for generating a paired dataset that combines multiple large models pretrained on different modalities: a large language model (GPT-3 [7]) and a text-to-image model (Stable Diffusion [51]). These two models capture complementary knowledge about language and images that can be combined to create paired training data for a task spanning both modalities.

Using our generated paired data, we train a conditional diffusion model that, given an input image and a text instruction for how to edit it, generates the edited image. Our model directly performs the image edit in the forward pass, and does not require any additional example images, full descriptions of the input/output images, or per-example fine-tuning. Despite being trained entirely on synthetic examples (i.e., both generated written instructions and generated

Tim Brooks, Alexander Holynski, and Alexei Efros
UC Berkeley, 2023

What does it do?

You give it an image

Then tell it to edit it in some way (the instruction)

It does it

[InstructPix2Pix - a Hugging Face Space by timbrooks]

"give him a Santa hat"

How does it differ from other image generation tools?

Unlike DreamBooth, InstructPix2Pix only needs one image

No need for text labels, captions, or descriptions of the input image

But how?

It’s a conditional diffusion model

It was trained on synthetic image-instruction-image triplets

One challenge: image consistency for ground truth

Limitations

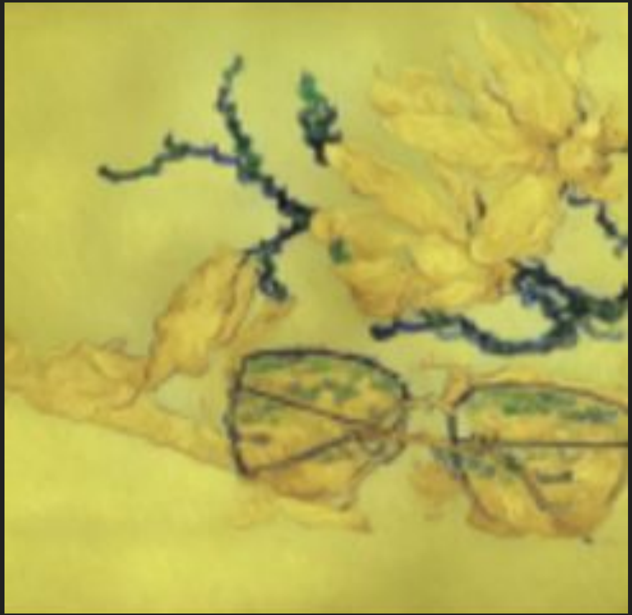
It won’t:

- shift objects,
- zoom in/out, or
- do things like “put two cups on the table and one on the chair.”

END OF RAÚL's SLIDES

Cross-Dataset Performance Evaluation: Dreambooth Images on InstructPix2Pix

Some things that went well…



"make it van gogh style."

"Convert to black and white."

Some things that didn't go well…

"move it to the beach."

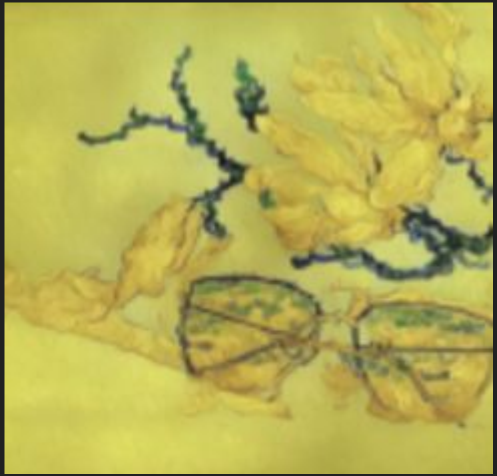
"make him wear a beanie."

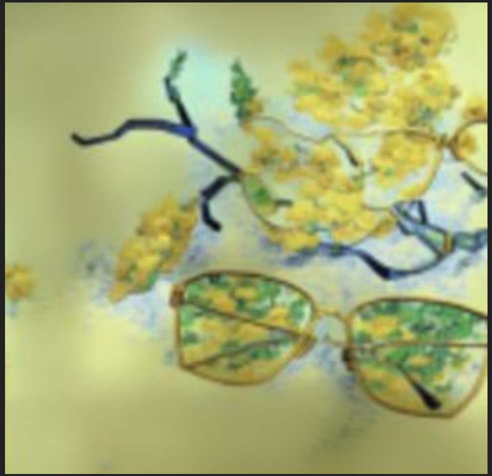
What’s the difference?

Global “filters” vs. specific objects

"make it van gogh style."

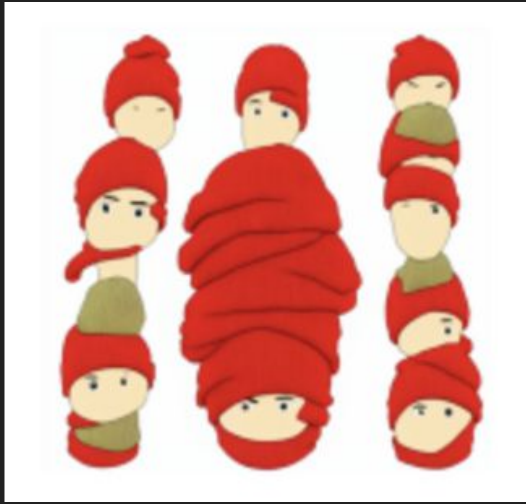
"make the sunglasses van gogh style."





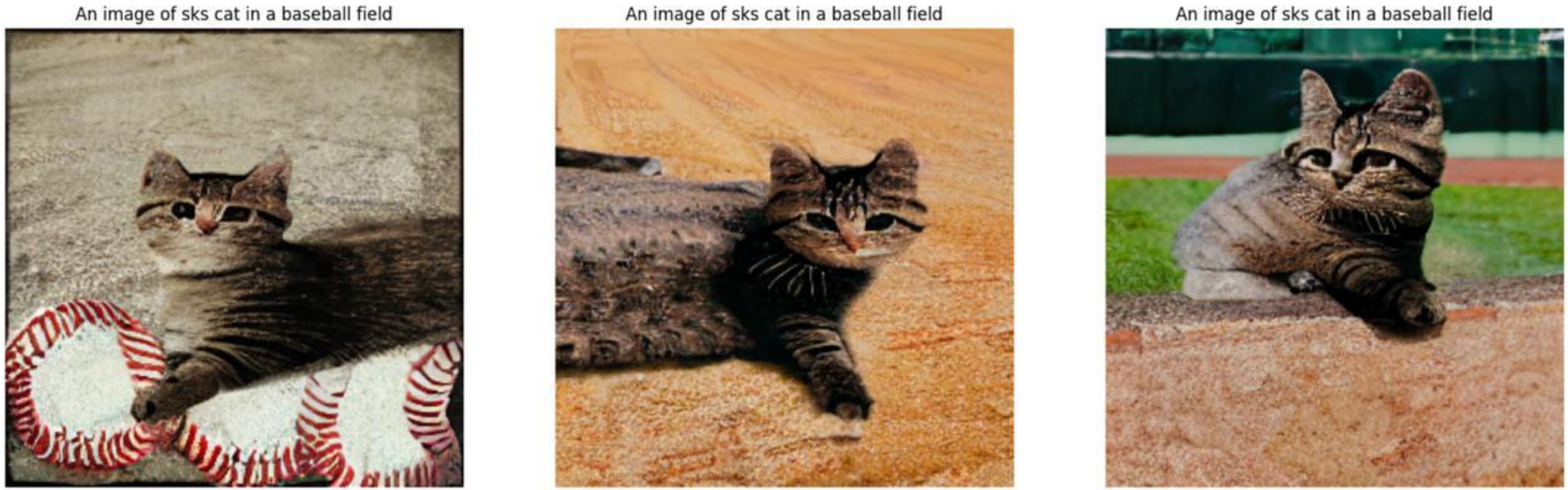
"make him wear a beanie."

"make the cartoon character wear a beanie."



Final Analysis: Dreambooth and InstructPix2Pix on Unseen Data

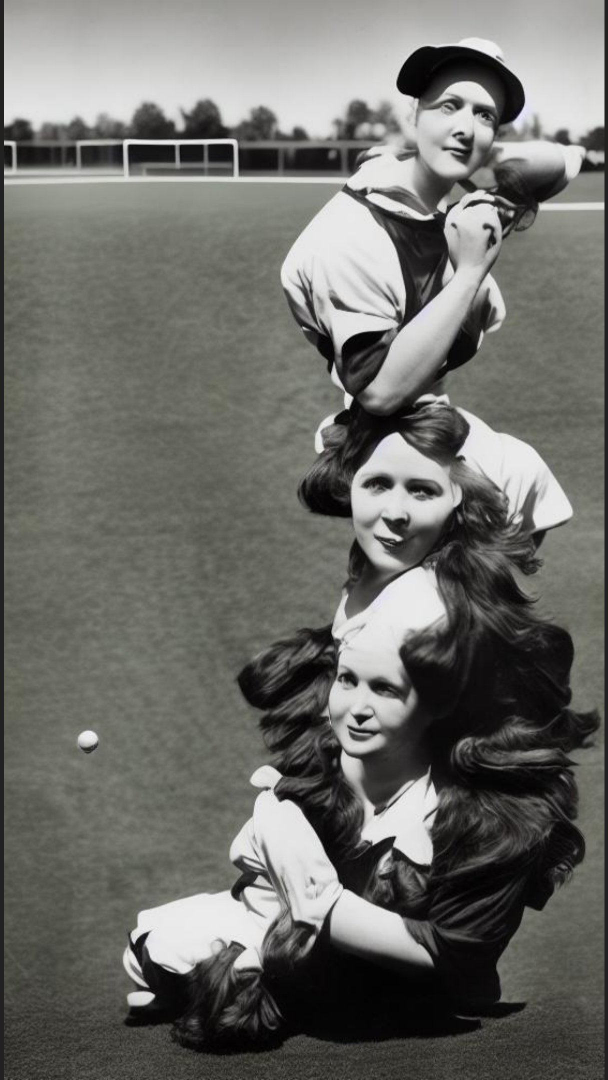
Dreambooth



“An image of sks cat in a baseball field”

InstructPix2Pix

“place her in a



baseball field”

“Put him at a

baseball game”

“Put the cat at a  baseball game”