WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Visual Instruction Inversion:

Image Editing via Visual Prompting

Thao Nguyen Yuheng Li Utkarsh Ojha Yong Jae Lee

Example

Before *After* *Test* *Output*

Text-guided Image Editing *is great...*

Input Image

Instruction

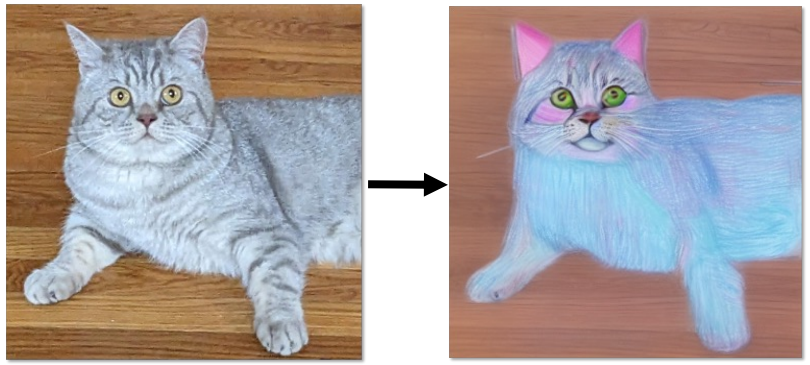
"Make it a drawing"

InstructPix2Pix

Output

Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

But what if the edit is **difficult** to describe in *text*?



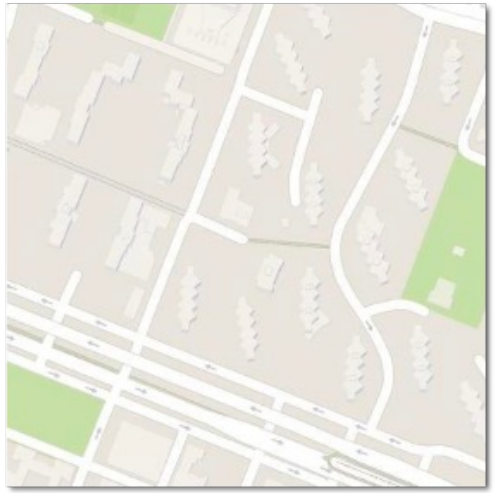
Before After

Input Instruction Output

Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

But what if the edit is **difficult** to describe in *text*?

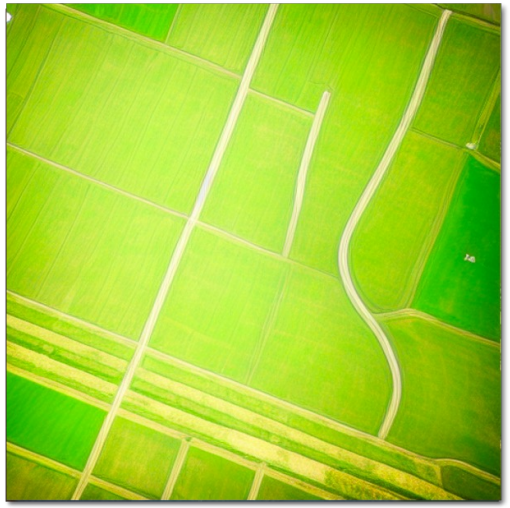
Before After



Input

"Turn it into a satellite photo"

Instruction



Output

Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

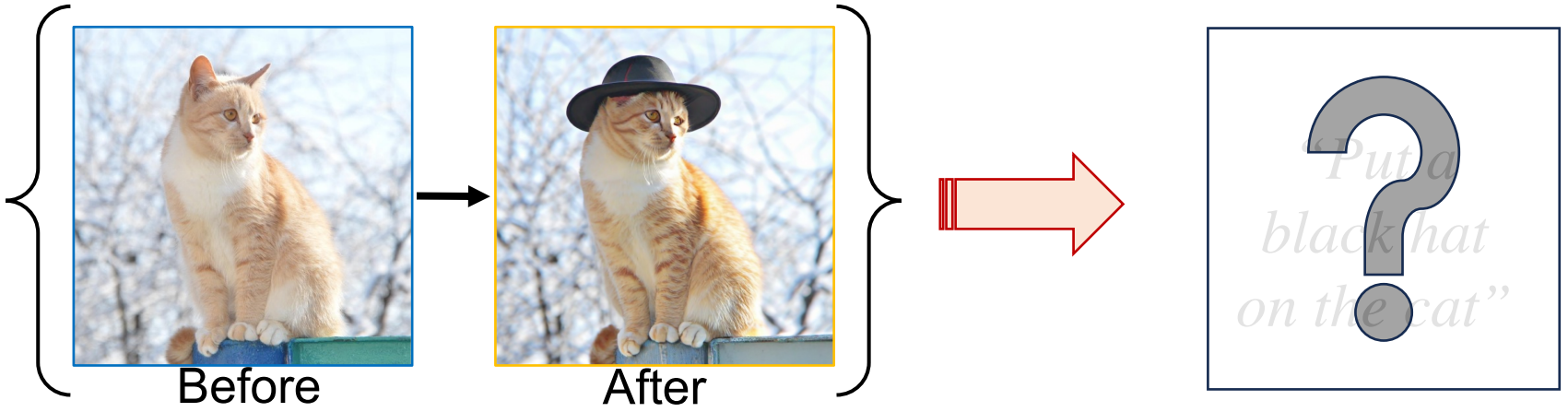
But what if the edit is **difficult** to describe in *text*?

Before After

Input Instruction Output

Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

How to invert visual prompts to edit instructions



Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

Inverting visual prompts to editing instructions (visii)

Before

editing instruction

InstructPix2Pix

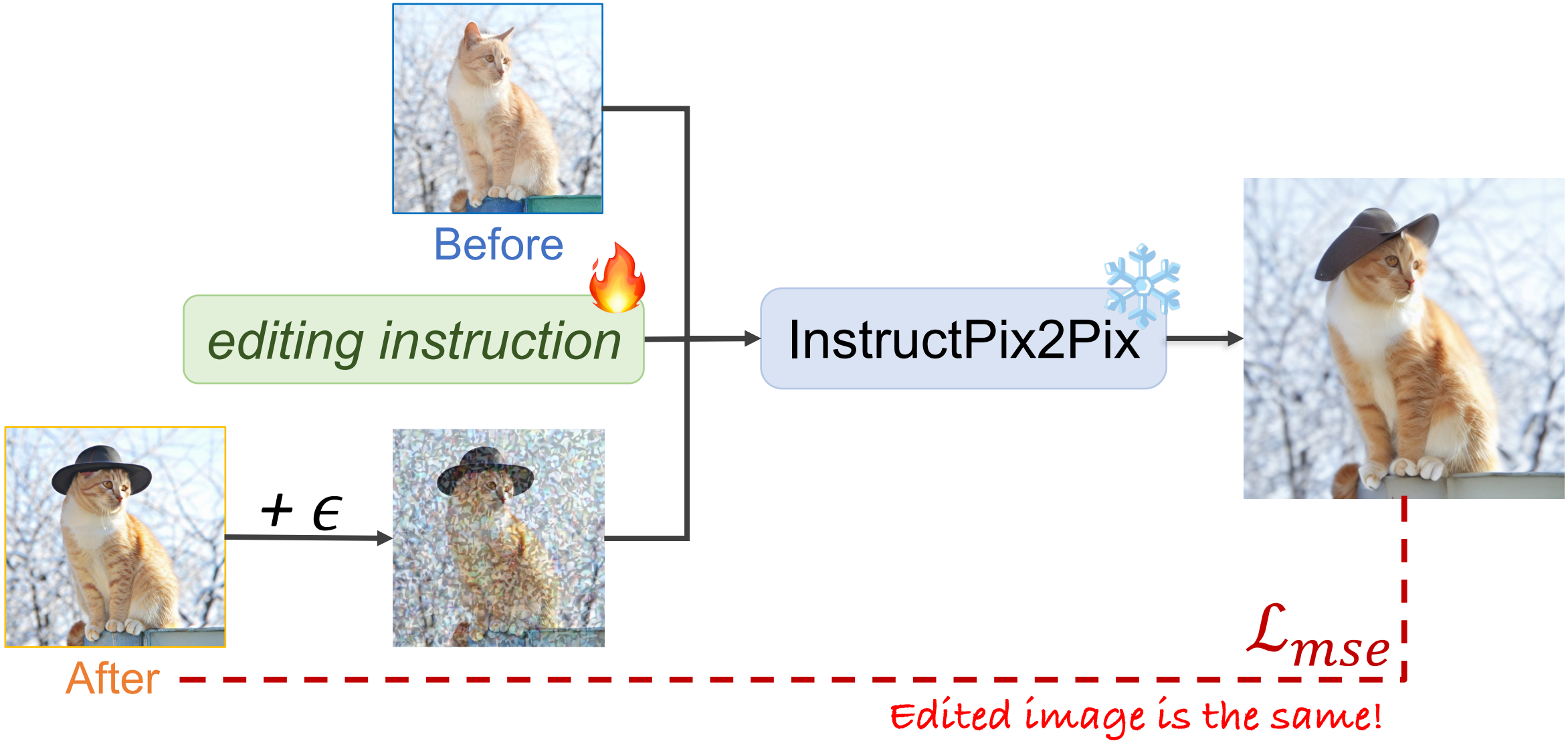
$+ \epsilon$

After

Idea: Optimize editing instructions!

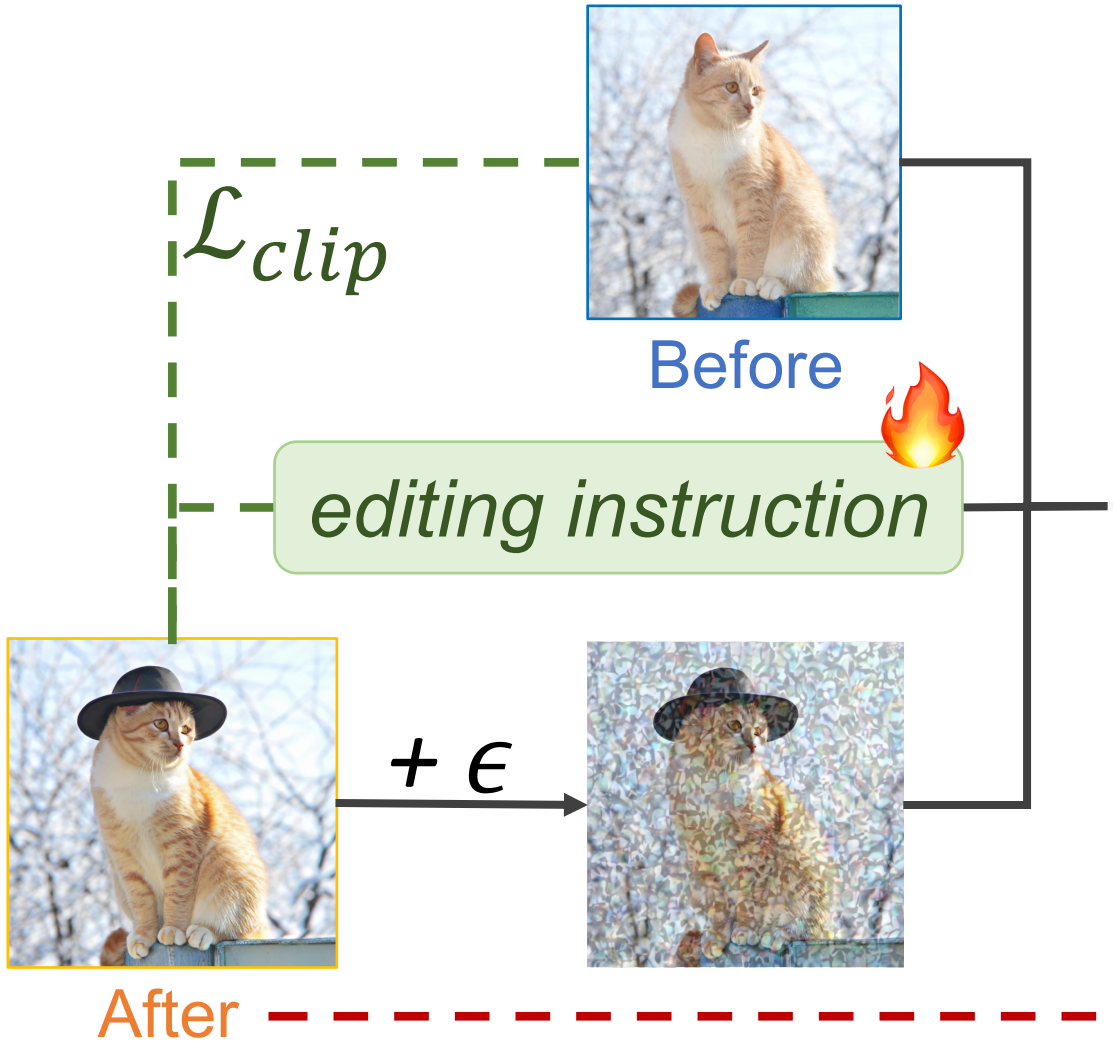
Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

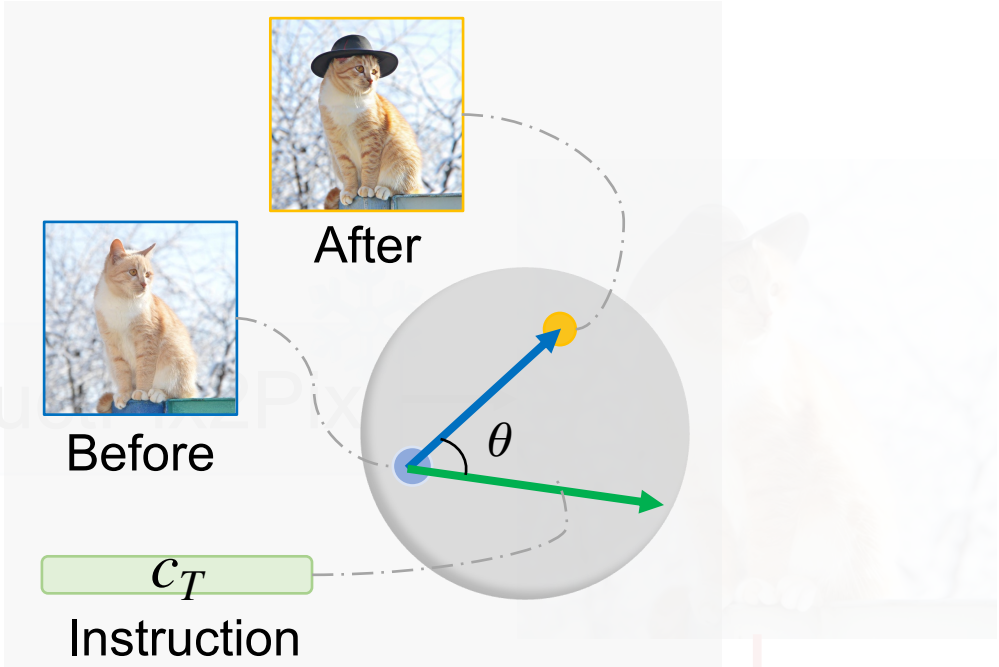
Inverting visual prompts to editing instructions (visii)



Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

Inverting visual prompts to editing instructions (visii)

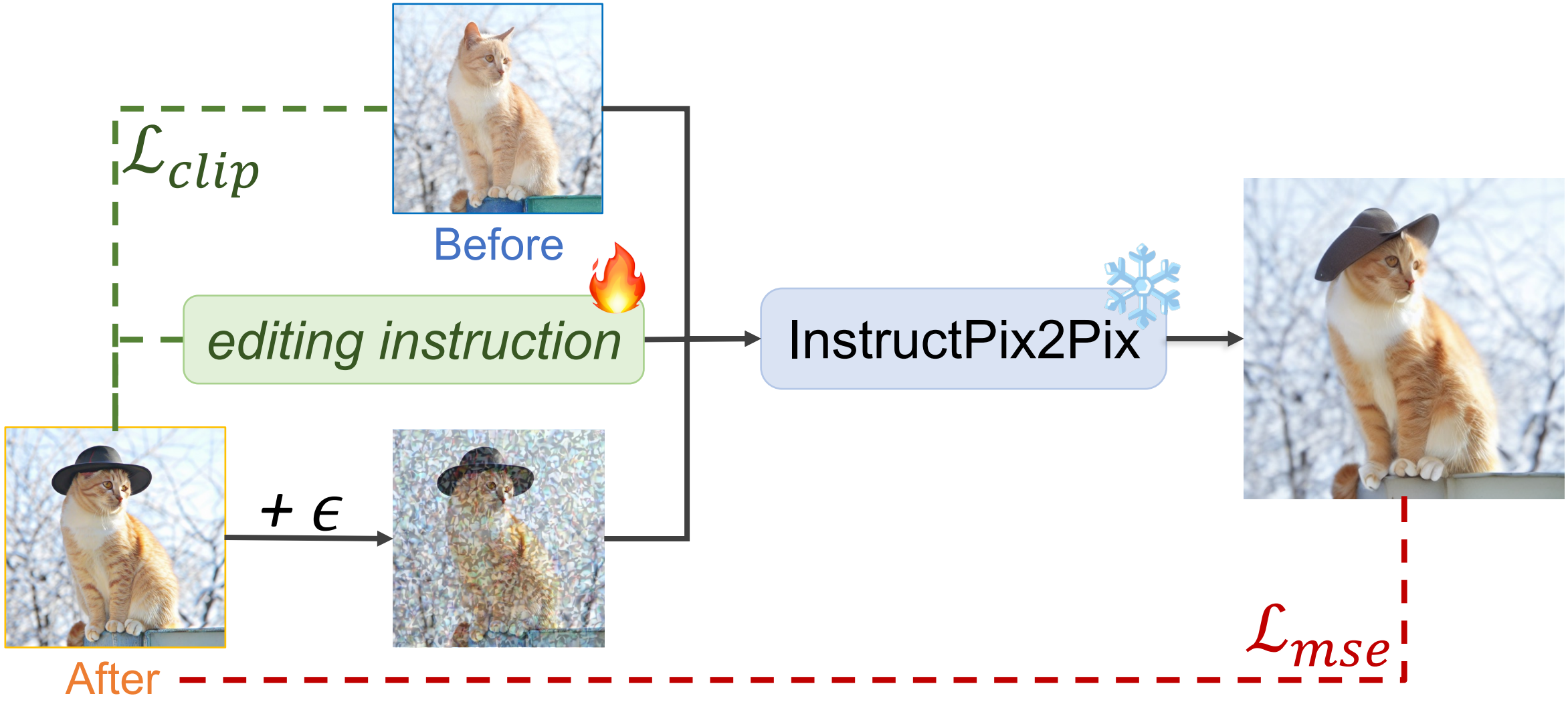




Instruction direction is aligned with before -> after direction

Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

Inverting visual prompts to editing instructions (visii)



Brooks et al., InstructPixPix: Learning to Follow Image Editing Instructions, CVPR 2023.

*“**Turn it into a painting**” can be many different edits*

Before

Test Images

*"**Turn it into a painting**" can be many different edits*

Before

After

Test Images

Results

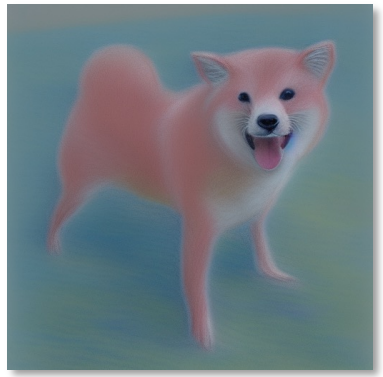
*"**Turn it into a painting**" can be many different edits*

Before

After

Test Images

Results



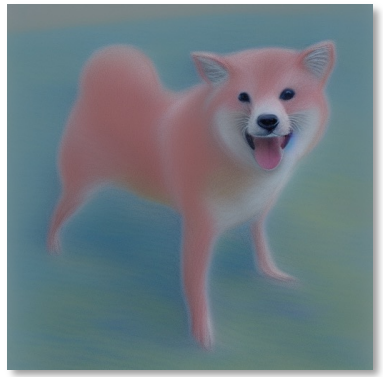
"Turn it into a painting" can be many different edits

Before

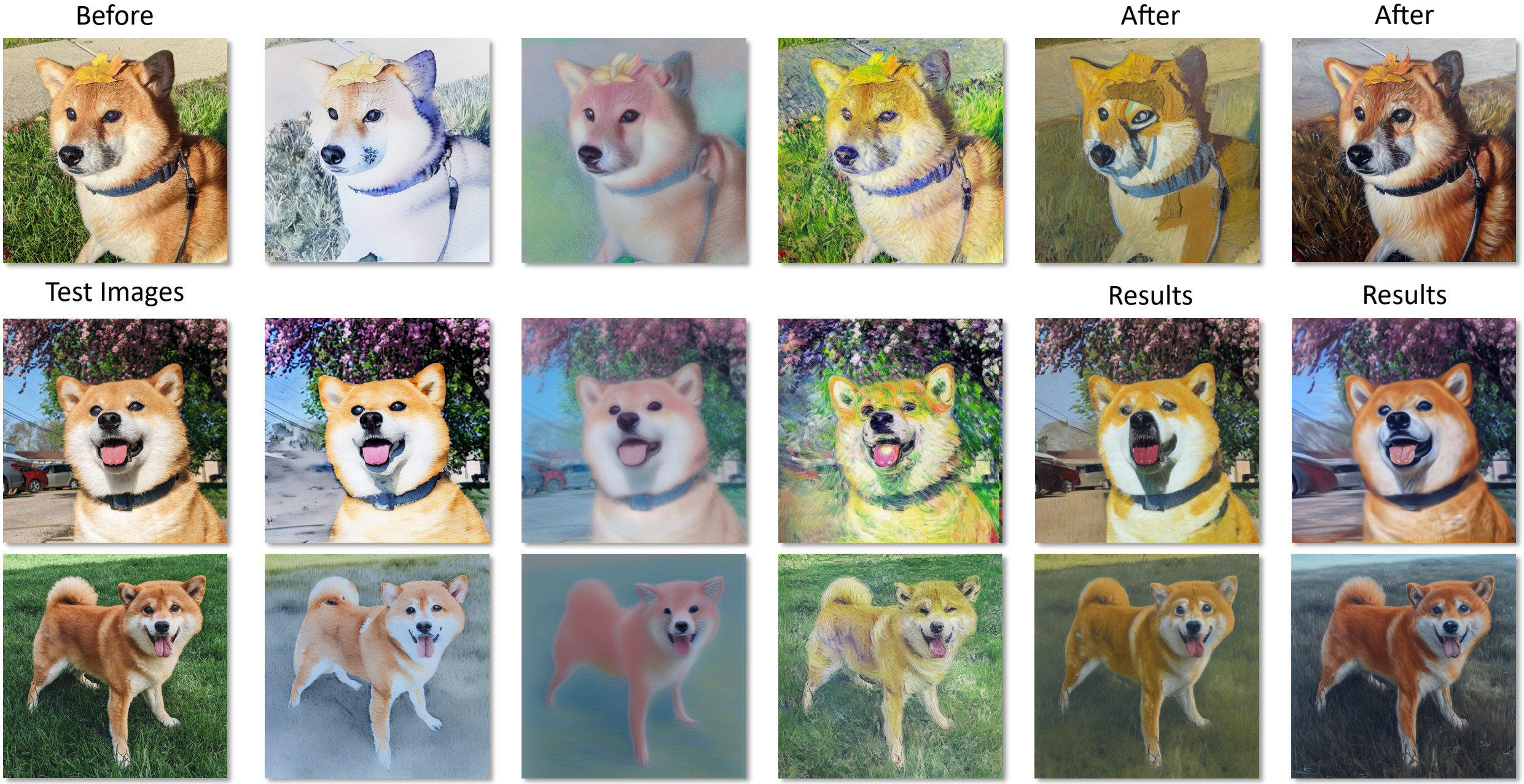
After

Test Images

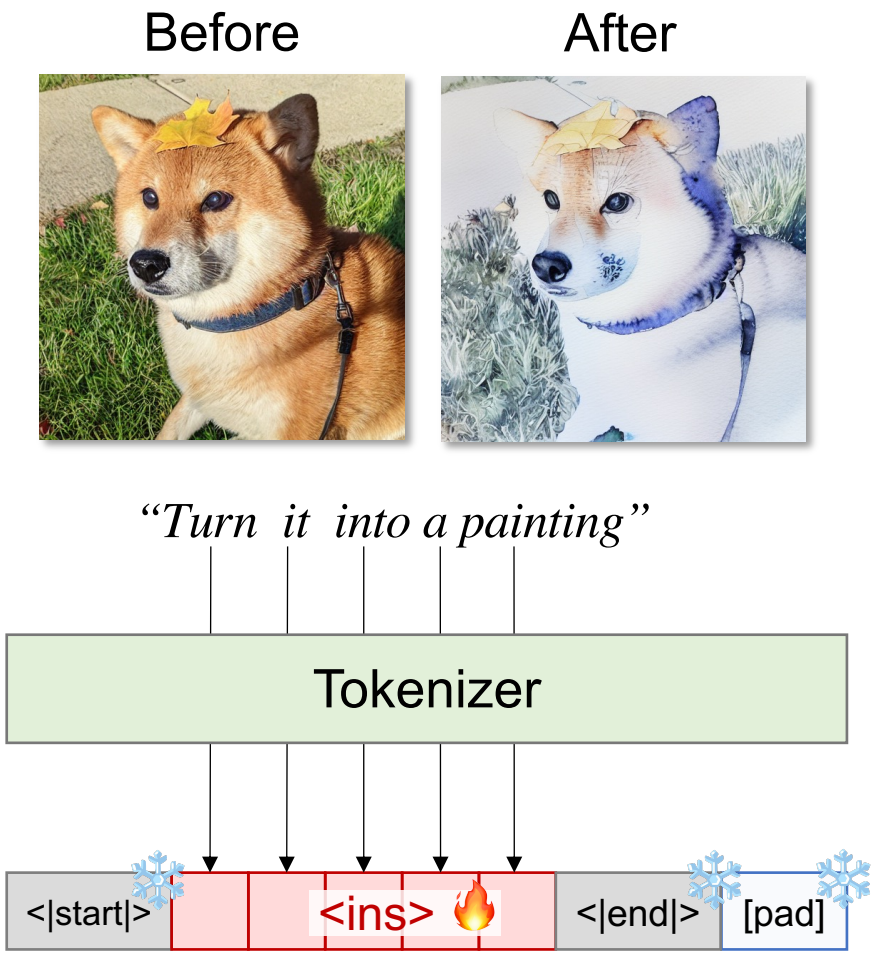
Results



"Turn it into a painting" can be many different edits



Hybrid instructions: Training



💡 **Idea**: Only optimize a fixed number of tokens!

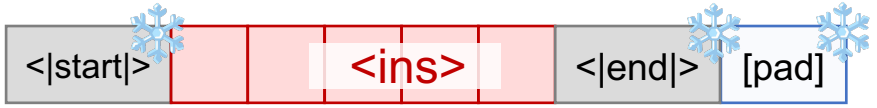
Hybrid instructions: Testing

Before

After

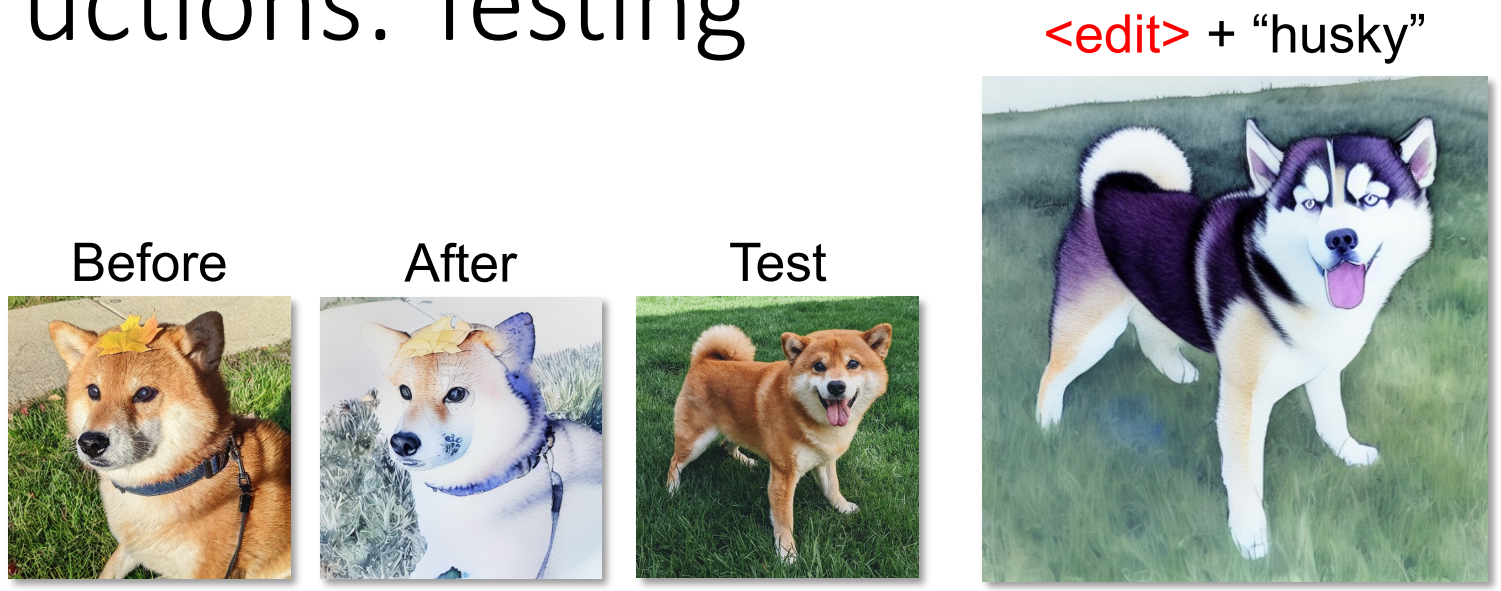
Test

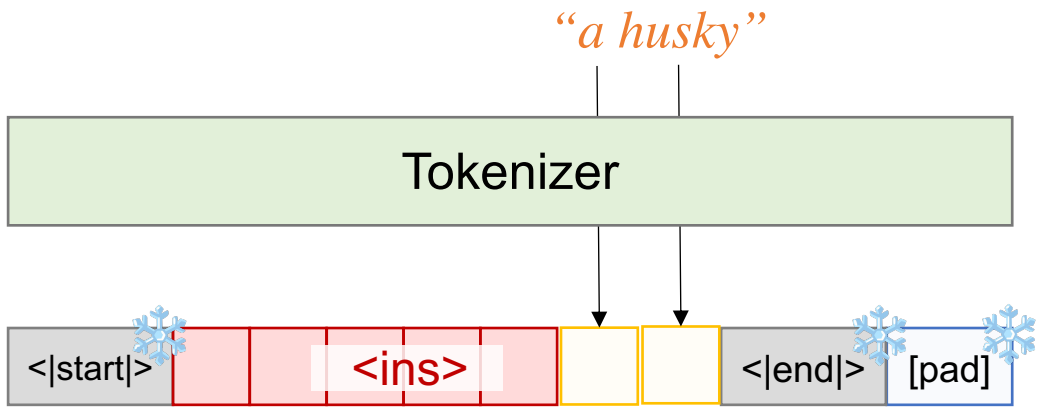
<edit>



💡 **Idea**: Can create hybrid instruction!

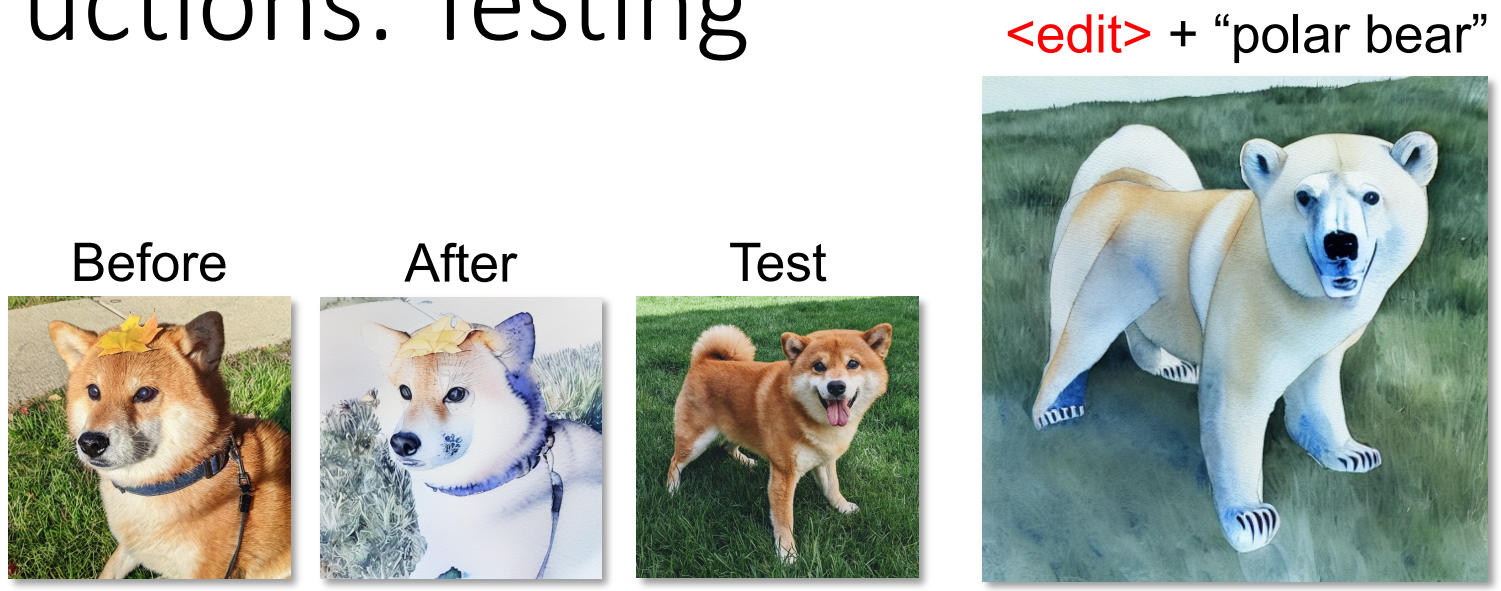
Hybrid instructions: Testing

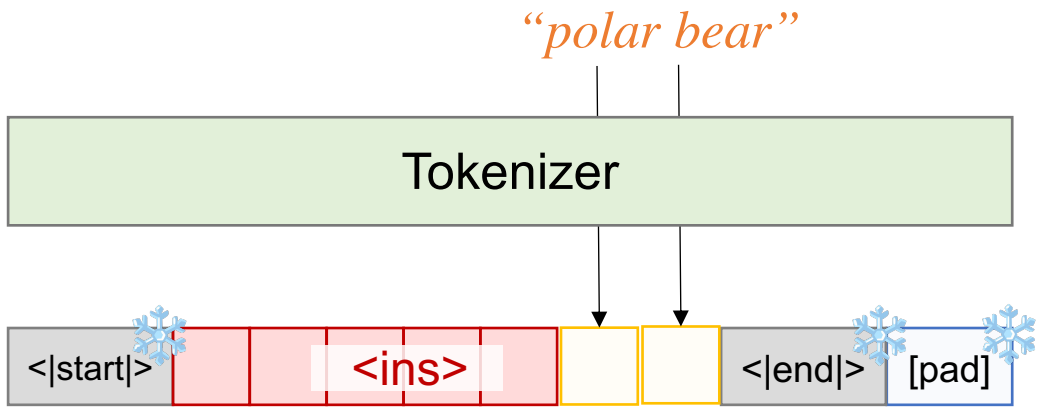




💡 **Idea**: Can create hybrid instruction!

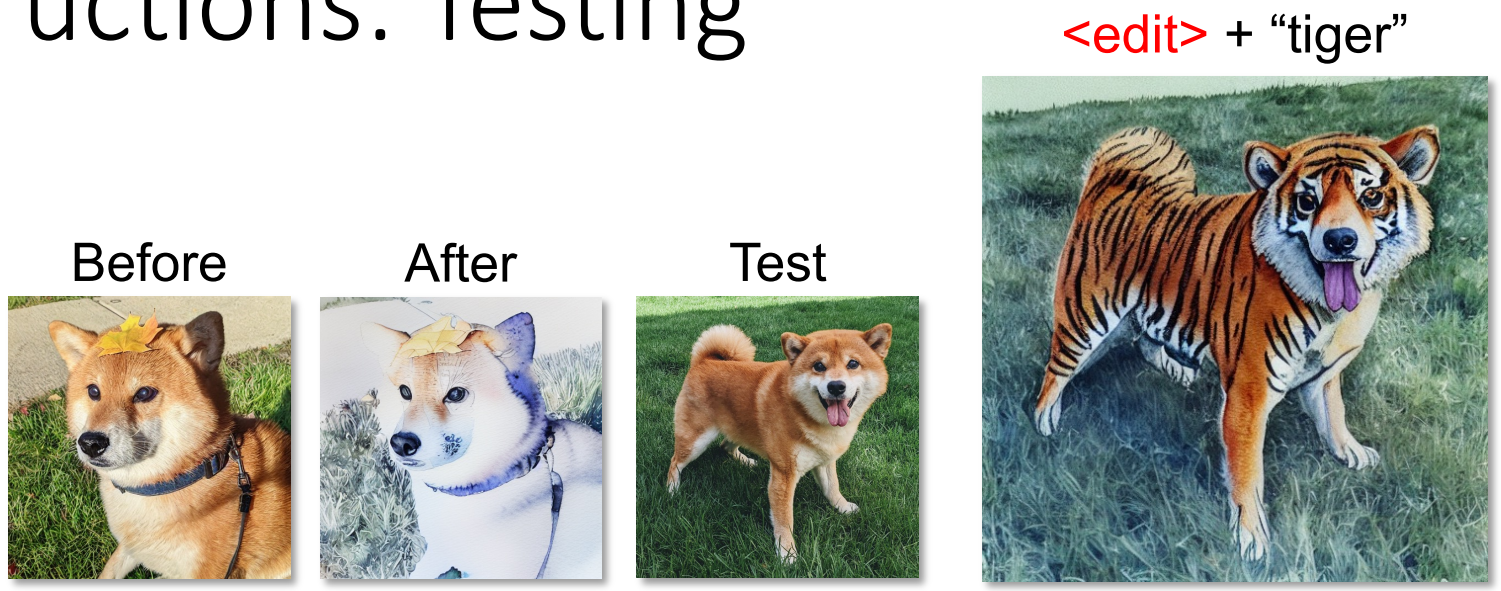
Hybrid instructions: Testing

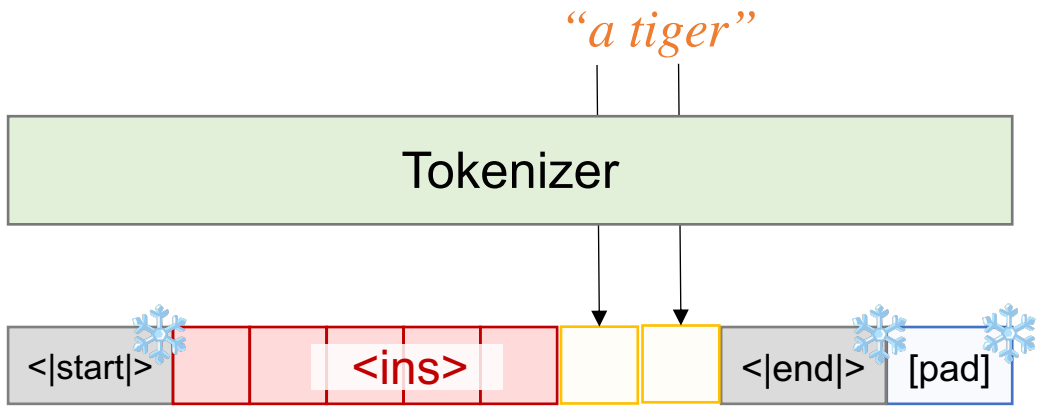




💡 **Idea**: Can create hybrid instruction!

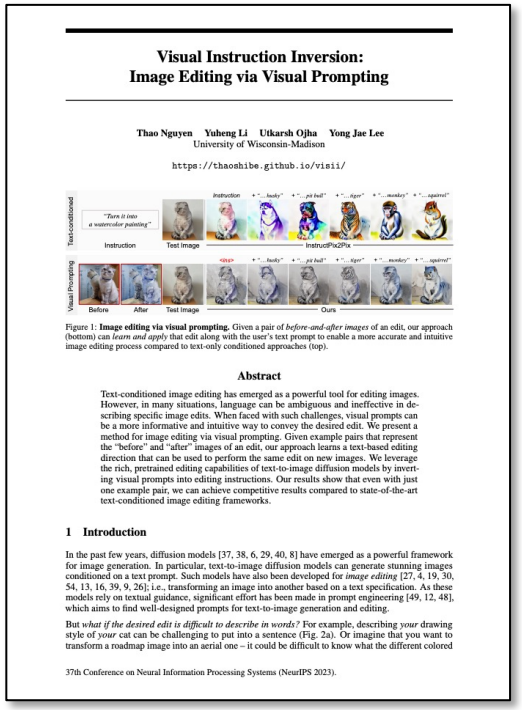
Hybrid instructions: Testing





💡 **Idea**: Can create hybrid instruction!

Visual Instruction Inversion:
Image Editing via Visual Prompting



Visual Instruction Inversion:
Image Editing via Visual Prompting

Thao Nguyen **Yuheng Li** **Utkarsh Ojha** **Yong Jae Lee**
University of Wisconsin-Madison

https://thaoshibe.github.io/visii/

Figure 1: **Image editing via visual prompting.** Given a pair of *before-and-after images* of an edit, our approach (bottom) can *learn and apply* that edit along with the user's text prompt to enable a more accurate and intuitive image editing process compared to text-only conditioned approaches (top).

Abstract

Text-conditioned image editing has emerged as a powerful tool for editing images. However, in many situations, language can be ambiguous and ineffective in describing specific image edits. When faced with such challenges, visual prompts can be a more informative and intuitive way to convey the desired edit. We present a method for image editing via visual prompting. Given example pairs that represent the "before" and "after" images of an edit, our approach learns a text-based editing direction that can be used to perform the same edit on new images. We leverage the rich, pretrained editing capabilities of text-to-image diffusion models by inverting visual prompts into editing instructions. Our results show that even with just one example pair, we can achieve competitive results compared to state-of-the-art text-conditioned image editing frameworks.

1 Introduction

In the past few years, diffusion models [37, 38, 6, 29, 40, 8] have emerged as a powerful framework for image generation. In particular, text-to-image diffusion models can generate stunning images conditioned on a text prompt. Such models have also been developed for *image editing* [27, 4, 19, 30, 54, 13, 16, 39, 9, 26]; i.e., transforming an image into another based on a text specification. As these models rely on textual guidance, significant effort has been made in prompt engineering [49, 12, 48], which aims to find well-designed prompts for text-to-image generation and editing.

But *what if the desired edit is difficult to describe in words?* For example, describing *your* drawing style of *your* cat can be challenging to put into a sentence (Fig. 2a). Or imagine that you want to transform a roadmap image into an aerial one – it could be difficult to know what the different colored

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A framework for
inverting visual prompts into editing instructions
for text-to-image diffusion models.

**** Project Page*: https://thaoshibe.github.io/visii/