GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos

Tomáš Souček[1], Dima Damen[2], Michael Wray[2], Ivan Laptev[3], Josef Šivic[1]

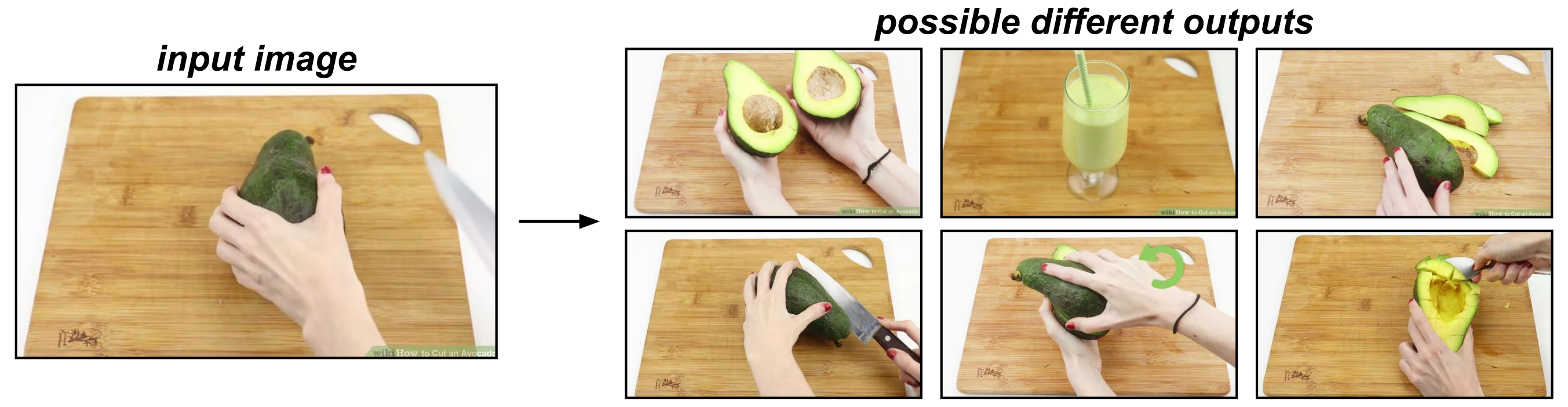
[1]Czech Technical University [2]University of Bristol [3]MBZUAI

CVPR JUNE 17-21, 2024 SEATTLE, WA

OVERVIEW

Goal

- Generate images showcasing **actions** exerted upon objects and the resulting **object state transformations** while preserving the scene



Motivation

- Learning goal-conditioned policies where goals are defined by images

Image taken from [7]

Challenges

- Obtaining paired training data is challenging
- Preservation of parts of the scene unrelated to the transformation (i.e., background and other objects)
- Introduction of new objects consistent with the transformation (e.g., a knife and hands for cutting)

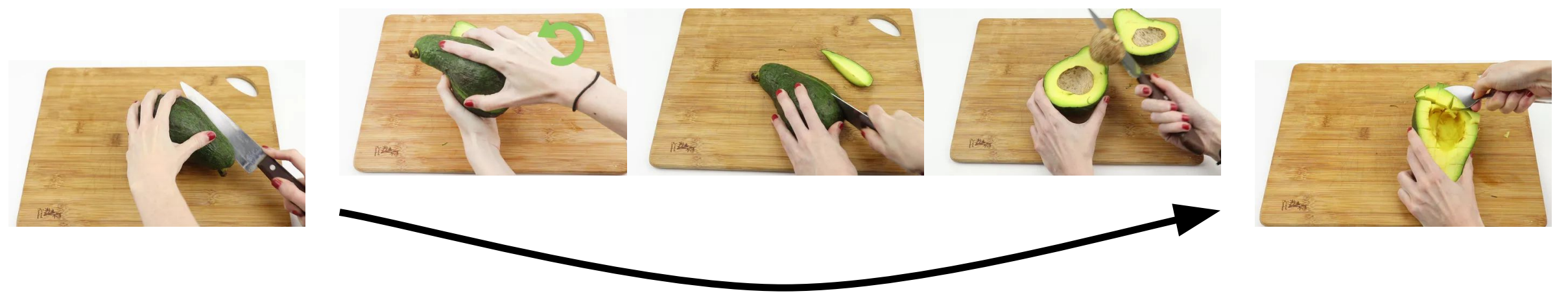
Contributions

- Dataset of 200K triplet images, mined from instructional HowTo videos, for training and evaluation
- GenHowTo: A text-conditioned generative model that produces the action image or the final image from an initial image of an action
- New quantitative evaluation, that uses classification to evaluate the state of generated images

DATASET CONSTRUCTION

1. INSTRUCTIONAL VIDEOS

- COIN & ChangeIt datasets
- 45k videos with (mostly) static background depicting object state changes



2. DETECT STATES & ACTIONS

- Self-supervised method [5] "discovers" **object states** and **actions** in the videos

3. CAPTION FRAMES

- Automatically caption [6] action and final state frames:

Action prompt $\mathcal{P}_{ac}$: ***a person cutting a fish on a cutting board***
Final state prompt $\mathcal{P}_{st}$: ***two pieces of fish on a wooden cutting board***

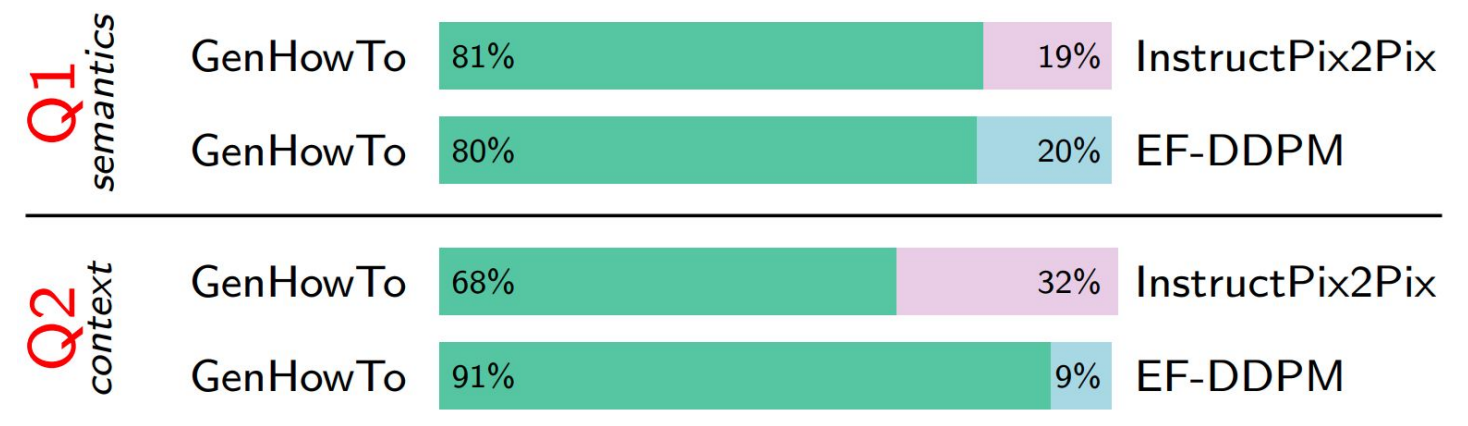
Input image $\mathcal{I}$ Action target $\mathcal{I}^*_{ac}$ Final state target $\mathcal{I}^*_{st}$

Action prompt $\mathcal{P}_{ac}$: ***a person slicing an apple on a cutting board***
Final state prompt $\mathcal{P}_{st}$: ***sliced apples on a cutting board next to a fork***

Input image $\mathcal{I}$ Action target $\mathcal{I}^*_{ac}$ Final state target $\mathcal{I}^*_{st}$

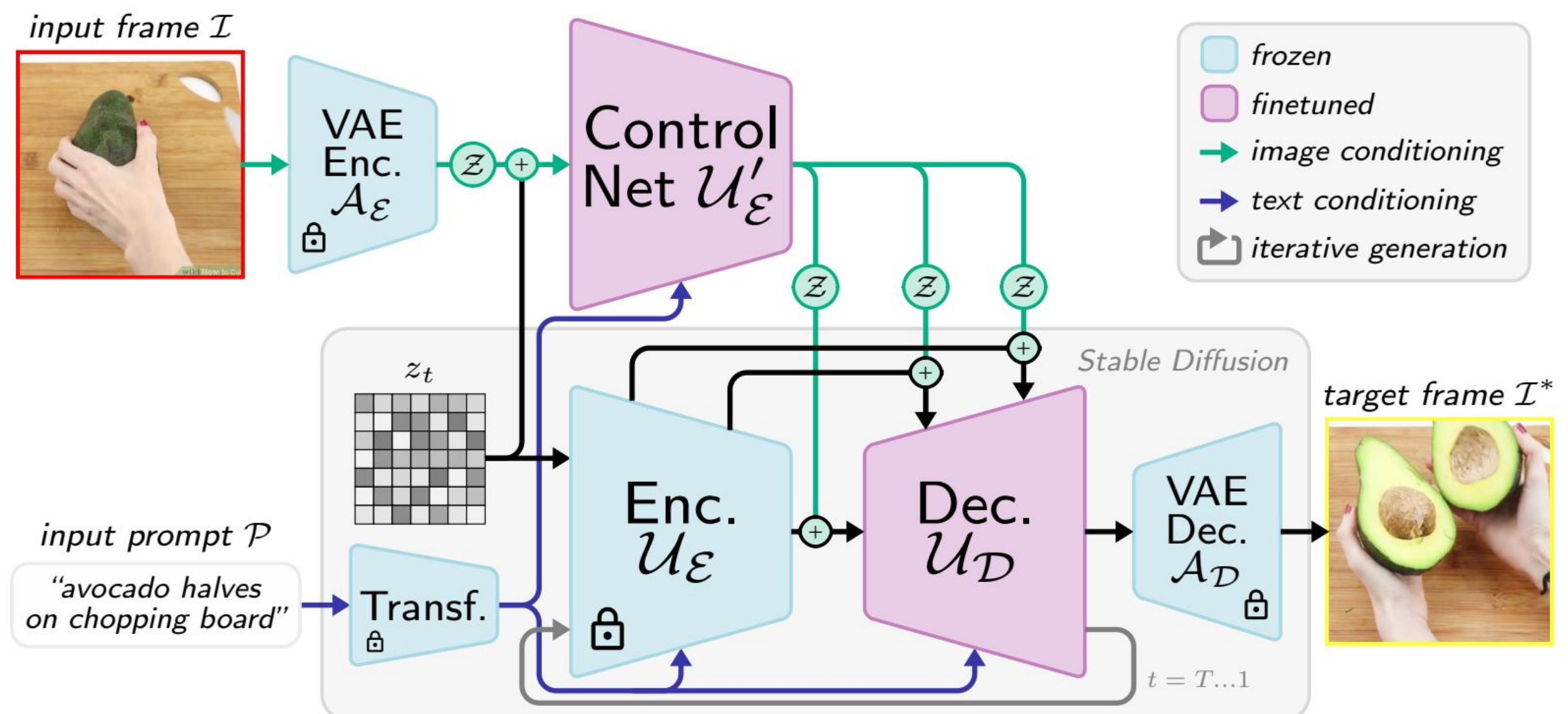
User study: 10 people, 2000 images

Q1: Which image better represents the final state described as of the same object as in the first image?
Q2: Which image better preserves the consistency of the scene?



MODEL

- **Model:** Stable Diffusion with ControlNet
- **Training data:**
  Input: initial state frame
  + action or final state prompt
  Output: action or final state frame


input frame $\mathcal{I}$
VAE Enc. $\mathcal{A}_\mathcal{E}$
$\mathcal{Z}$
Control Net $\mathcal{U}'_\mathcal{E}$
frozen
finetuned
image conditioning
text conditioning
iterative generation
z_t
Stable Diffusion
Enc. $\mathcal{U}_\mathcal{E}$
Dec. $\mathcal{U}_\mathcal{D}$
VAE Dec. $\mathcal{A}_\mathcal{D}$
target frame $\mathcal{I}^*$
input prompt $\mathcal{P}$
"avocado halves on chopping board"
Transf.
$t = T...1$

QUANTITATIVE RESULTS

New protocol to evaluate image transformation based on classification

- **Goal:** learn a classifier to discriminate initial, action, and final state images
- **Test set:** real initial, real action, and real final state images
- **Train set:** real initial state images, generated action and generated final state images

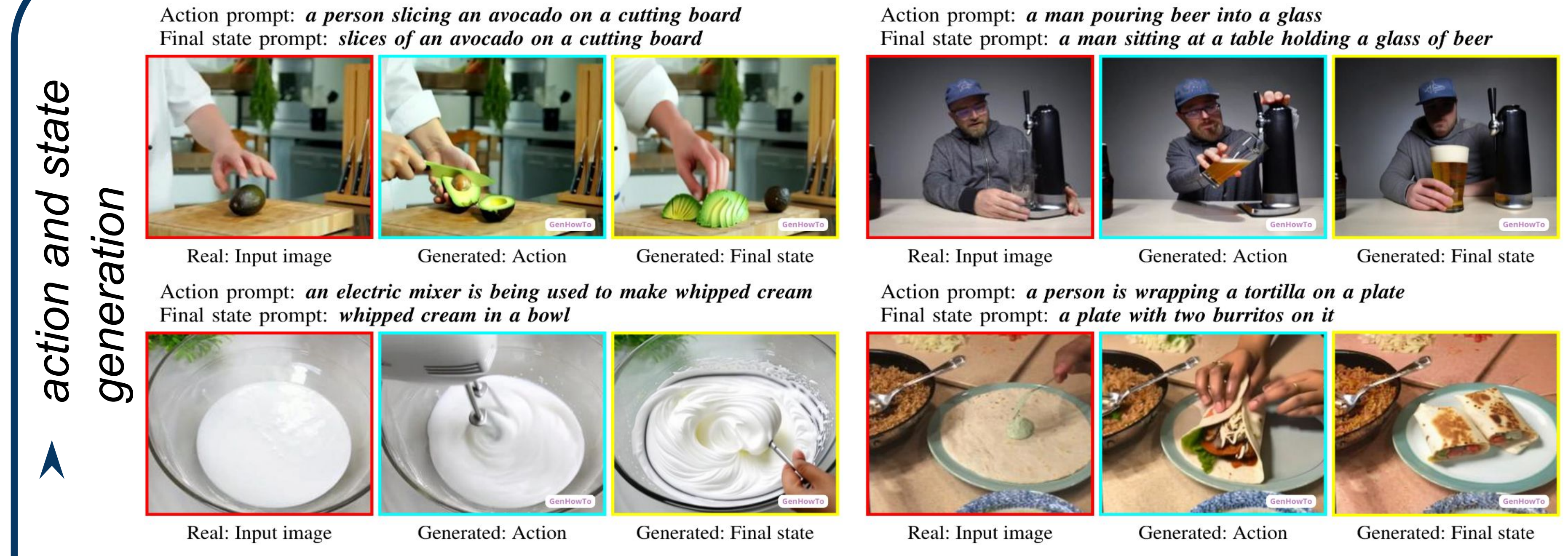
Method	Acc_{ac}	Acc_{st}
test set categories unseen during training		
Stable Diffusion [1]	0.51	0.50
Edit Friendly DDPM [2]	0.60	0.61
InstructPix2Pix [3]	0.55	0.63
CLIP (manual prompts) [4]	0.52	0.62
Ours (GenHowTo)	**0.66**	**0.74**
test set categories seen during training		
Edit Friendly DDPM [2]	0.69	0.80
Ours (GenHowTo)	**0.77**	**0.88**
Real images	0.96	0.97

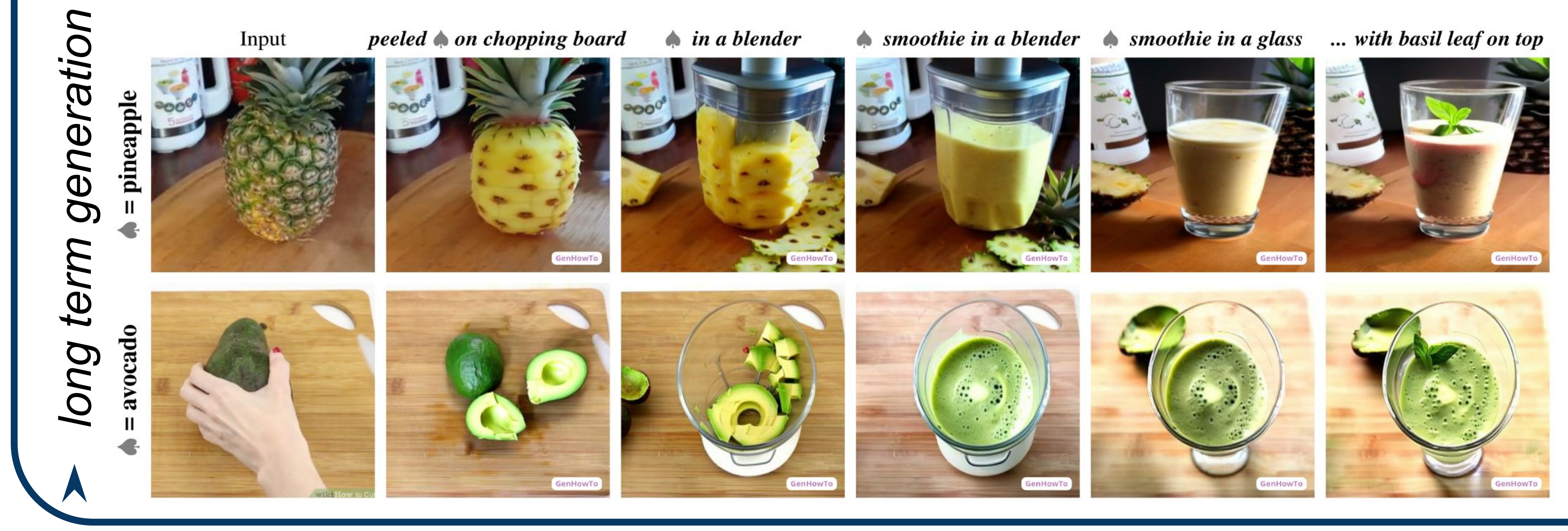
Ablations

- Using image captions and temporal frame detection [5] is important!

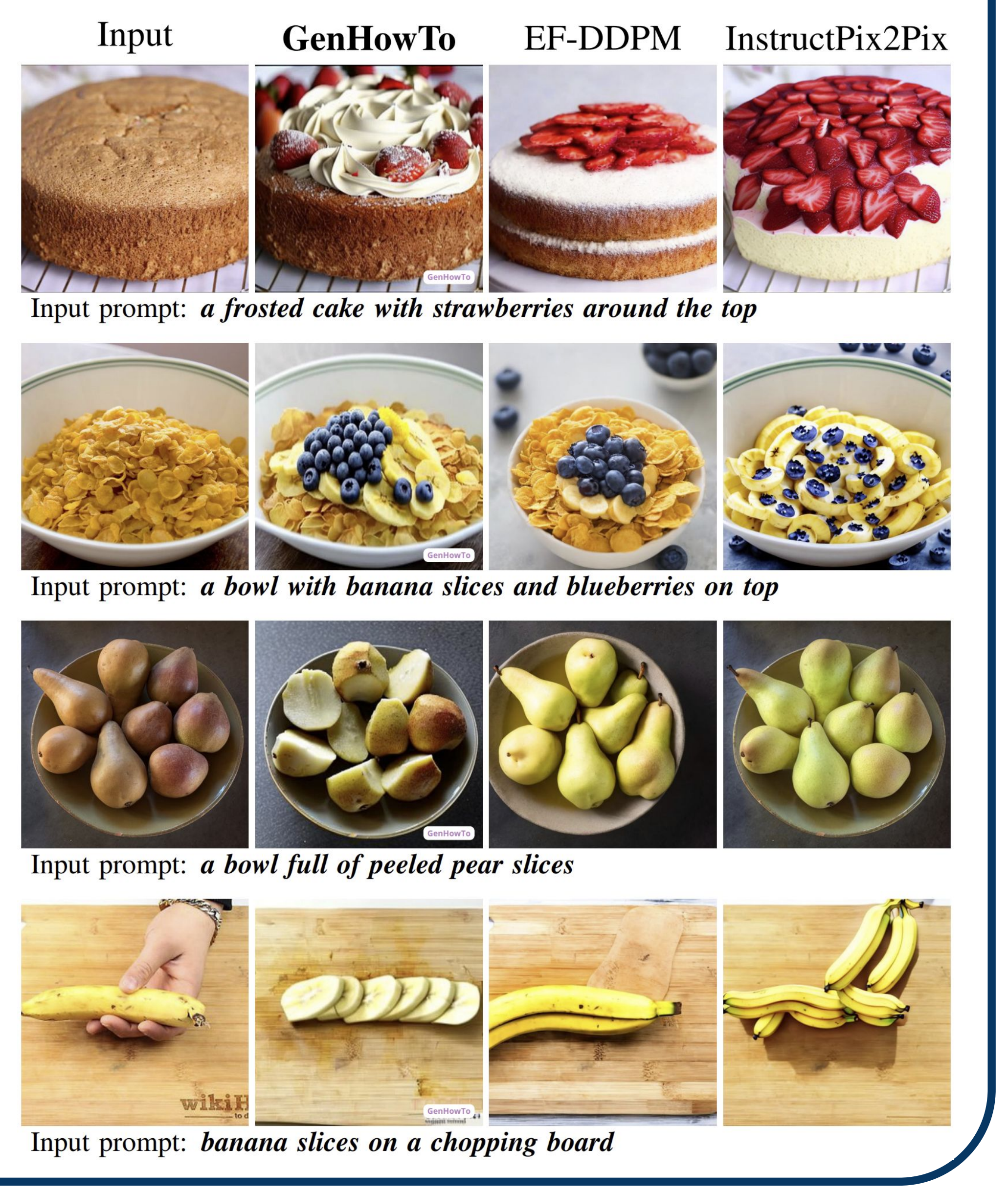
Method	Acc_{st}
Ours	0.74
Narrations (ASR) instead of captions	0.60
Uniformly sampled frames instead of [5]	0.67

QUALITATIVE RESULTS


action and state generation
Action prompt: a person slicing an avocado on a cutting board
Final state prompt: slices of an avocado on a cutting board
Real: Input image
Generated: Action
Generated: Final state
Action prompt: a man pouring beer into a glass
Final state prompt: a man sitting at a table holding a glass of beer
Real: Input image
Generated: Action
Generated: Final state
Action prompt: an electric mixer is being used to make whipped cream
Final state prompt: whipped cream in a bowl
Real: Input image
Generated: Action
Generated: Final state
Action prompt: a person is wrapping a tortilla on a plate
Final state prompt: a plate with two burritos on it
Real: Input image
Generated: Action
Generated: Final state
GenHowTo



- *comparison with related work*


Input
GenHowTo
EF-DDPM
InstructPix2Pix
Input prompt: a frosted cake with strawberries around the top
Input prompt: a bowl with banana slices and blueberries on top
Input prompt: a bowl full of peeled pear slices
Input prompt: banana slices on a chopping board
GenHowTo

Project page code & trained models

soczech.github.io/genhowto

[1] Rombach et al. High-resolution image synthesis with latent diffusion models. CVPR, 2022.

[2] Huberman-Spiegelglas et al. An edit friendly ddpm noise space: Inversion and manipulations. 2023.

[3] Brooks et al. InstructPix2Pix: Learning to follow image editing instructions. CVPR, 2023.

[4] Radford et al. Learning transferable visual models from natural language supervision. ICML, 2021.

[5] Souček et al. Multi-task learning of object states and state-modifying actions from web videos. TPAMI, 2024.

[6] BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML, 2023.

[7] E. Heiden et al. Disect: A differentiable simulation engine for autonomous robotic cutting. Robotics: Science and Systems, 2021.