LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning

ECCV 2024, Oral

Bolin Lai

Xiaoliang Dai

Lawrence Chen

Guan Pang

James M. Rehg

Miao Liu

Meta

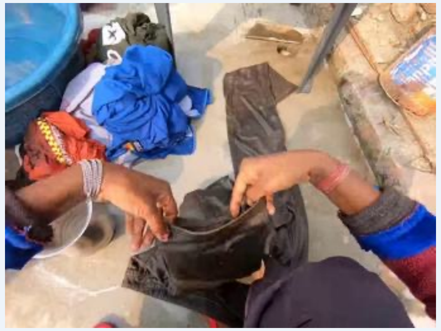
Georgia Tech

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Motivation

User Prompt

“Can you provide instructions on how to ***wash the trouser with the brush*** *in my current situation?”*



Egocentric Action Frame Generation

*“**1. Check the Care Label:** Check the care label on your trousers for any specific washing instructions.*

2. Prep the Area: *Find a clean and flat surface to work on, such as a table or countertop.*

***3. Spot Check Stains:**”*

⋮

(a) LLM Response

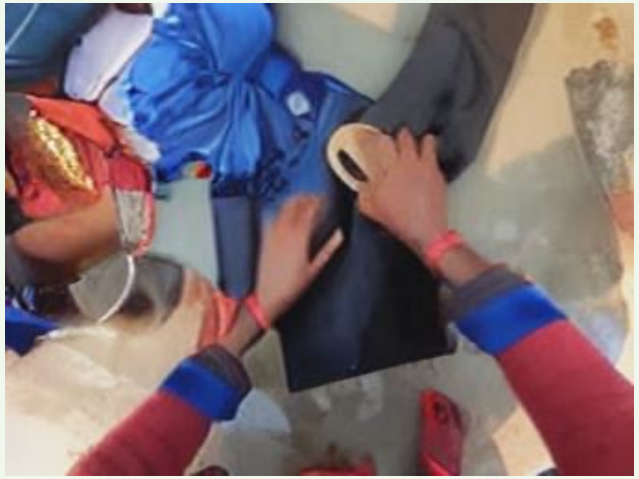
“1. You should submerge the trouser in the water.

2. Use the brush to scrub the trouser, focusing on any stains or areas that may require extra attention.

3. Once the trouser is clean, you should rinse it.”

⋮

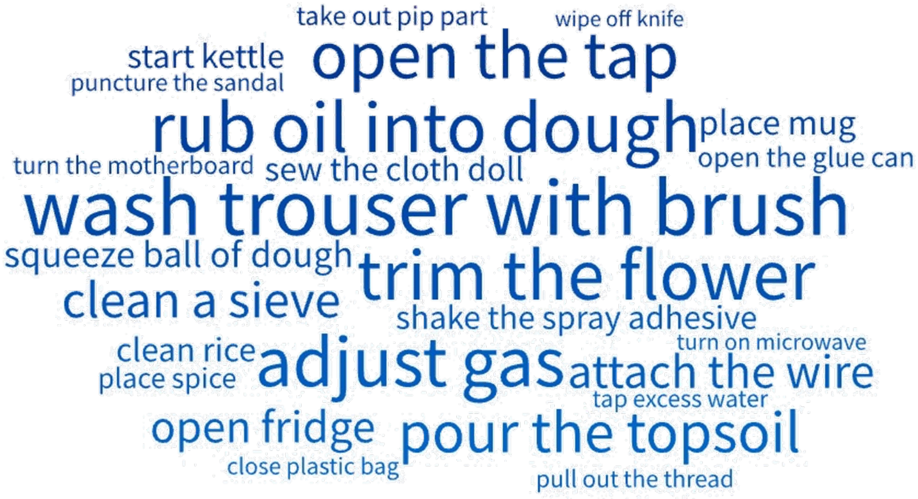
(b) Visual LLM Response



(c) Our model (LEGO) Response

Challenges

- Action labels are short of necessary details for action frame generation.



Action labels in existing egocentric datasets

- The off-the-shelf diffusion models are limited in action understanding due to domain gap.

"An armchair that looks like an apple"

"A photo of a maine coon, a type of pet"

"A dog rolling in the snow at sunset"

Data in diffusion model domain

"Return an egg in the tray"

"Repair a piston"

"Put the cement on the concrete"

Data in egocentric action domain

Challenges

- Action labels are short of necessary details for action frame generation.
- The off-the-shelf diffusion models are limited in action understanding due to domain gap.

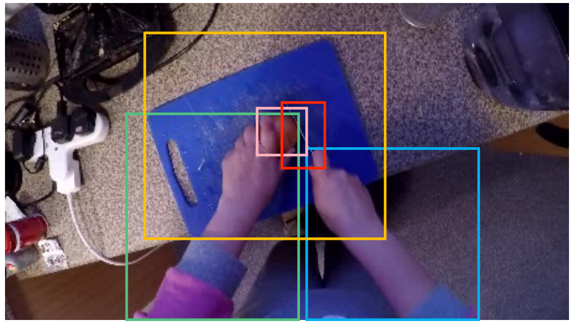
Our Solution:

Enriching the action labels with LLM via visual instruction tuning.

Leveraging finetuned LLM embeddings to improve egocentric action frame generation.

Data Curation

Example for In-context Learning



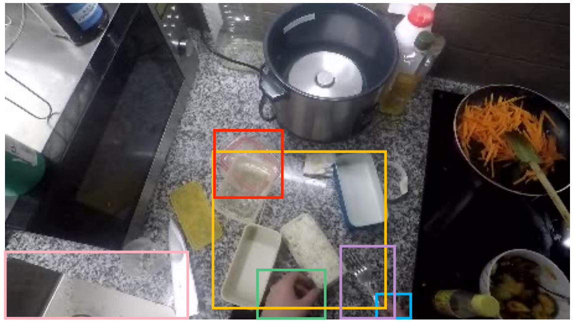
Input Query

Action Label: *"chop end of onion"*

Bounding Boxes: *"left hand-[0.203, 0.368, 0.499, 1.000], right hand-[0.530, 0.527, 0.797, 1.000], chopping board-[0.248, 0.156, 0.644, 0.750], onion-[0.462, 0.377, 0.509, 0.506], knife-[0.484, 0.343, 0.546, 0.586]"*

Action Description: *"The person presses the onion on the chopping board with the left hand and then cuts off the end of the onion with a knife in the right hand."*

Query and Response



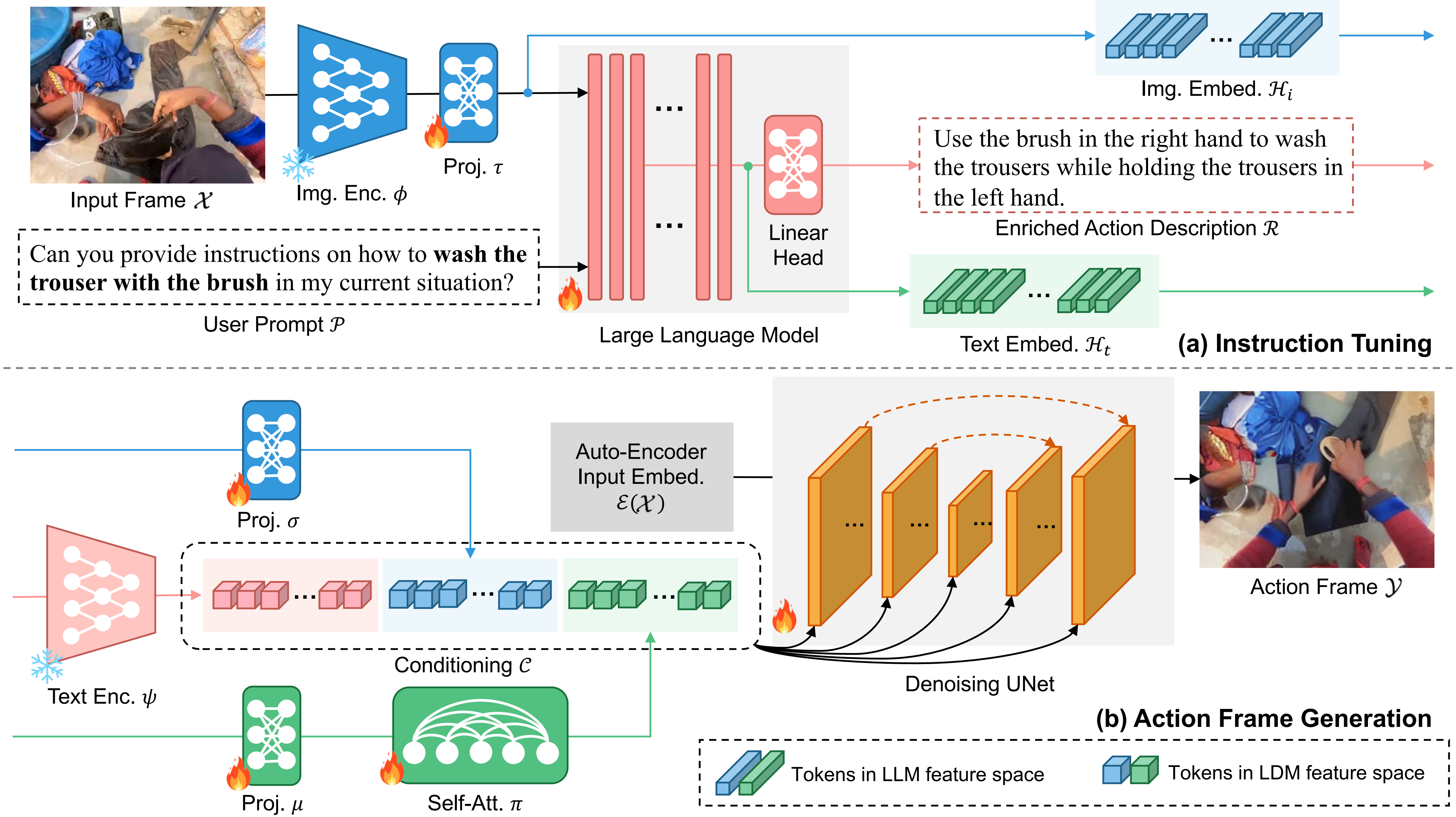
Input Query

Action Label: *"squidge into lunch box"*

Bounding Boxes: *"left hand-[0.427, 0.799, 0.591, 1.00], right hand-[0.654, 0.949, 0.707, 1.00], lunch box-[0.390, 0.512, 0.696, 0.993], sink-[0.00, 0.802, 0.315, 1.00], container lid-[0.403, 0.454, 0.513, 0.635], fork-[0.589, 0.798, 0.672, 1.00]"*

GPT-3.5 Response: *"The person uses their left hand to hold a lunch box, while their right hand uses a fork to squidge food into the lunch box, which is placed on the sink with the container lid nearby."*

Method



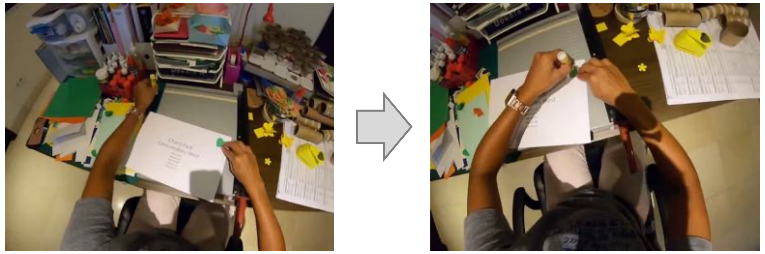
Can you provide instructions on how to **wash the trouser with the brush** in my current situation?

Use the brush in the right hand to wash the trousers while holding the trousers in the left hand.

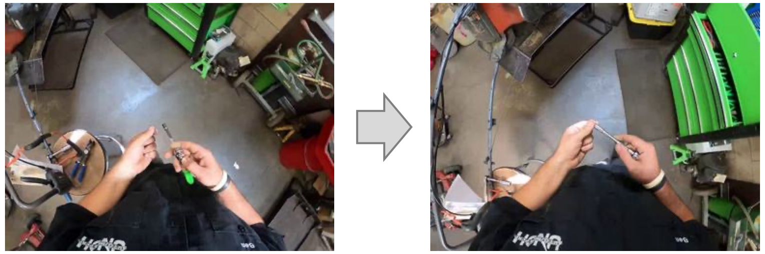
Input Frame $\mathcal{X}$
Img. Enc. ϕ
Proj. τ
Img. Embed. $\mathcal{H}_i$
Linear Head
Enriched Action Description $\mathcal{R}$
User Prompt $\mathcal{P}$
Large Language Model
Text Embed. $\mathcal{H}_t$
(a) Instruction Tuning
Proj. σ
Auto-Encoder Input Embed. $\mathcal{E}(\mathcal{X})$
Action Frame $\mathcal{Y}$
Text Enc. ψ
Conditioning $\mathcal{C}$
Denoising UNet
(b) Action Frame Generation
Proj. μ
Self-Att. π
Tokens in LLM feature space
Tokens in LDM feature space

Experiments – Datasets

Ego4D



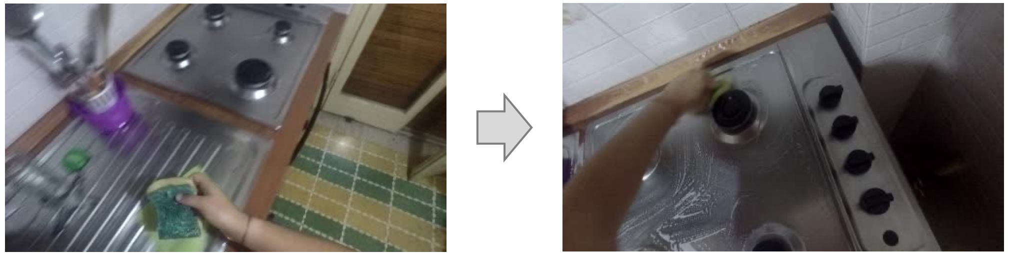
“apply glue on the manila paper”



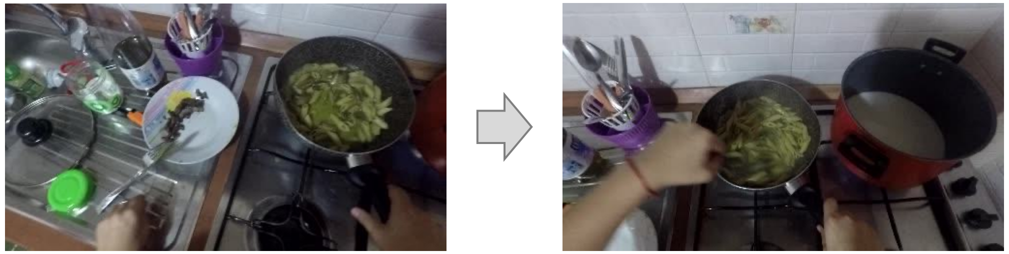
“fix a screw in screw driver”

- Include various actions in multiple scenarios.
- Use the pre-defined PRE-15 and PNR frames as input and target frames respectively.
- 85521 training samples, 9931 test samples

Epic-Kitchens



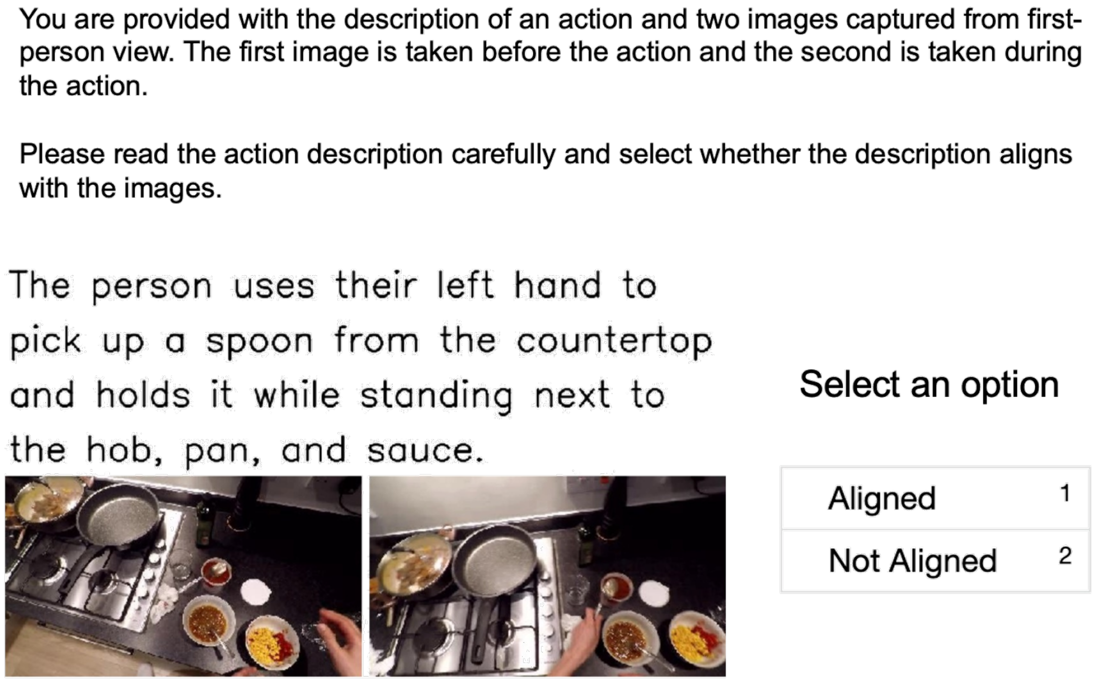
“wash cooktop”



“stir pasta”

- Include various actions in kitchens.
- Empirically define PRE-15 and PNR by ourselves.
- 61841 training samples, 8893 test samples

Experiments – Quality of Action Descriptions



You are provided with the description of an action and two images captured from first-person view. The first image is taken before the action and the second is taken during the action.

Please read the action description carefully and select whether the description aligns with the images.

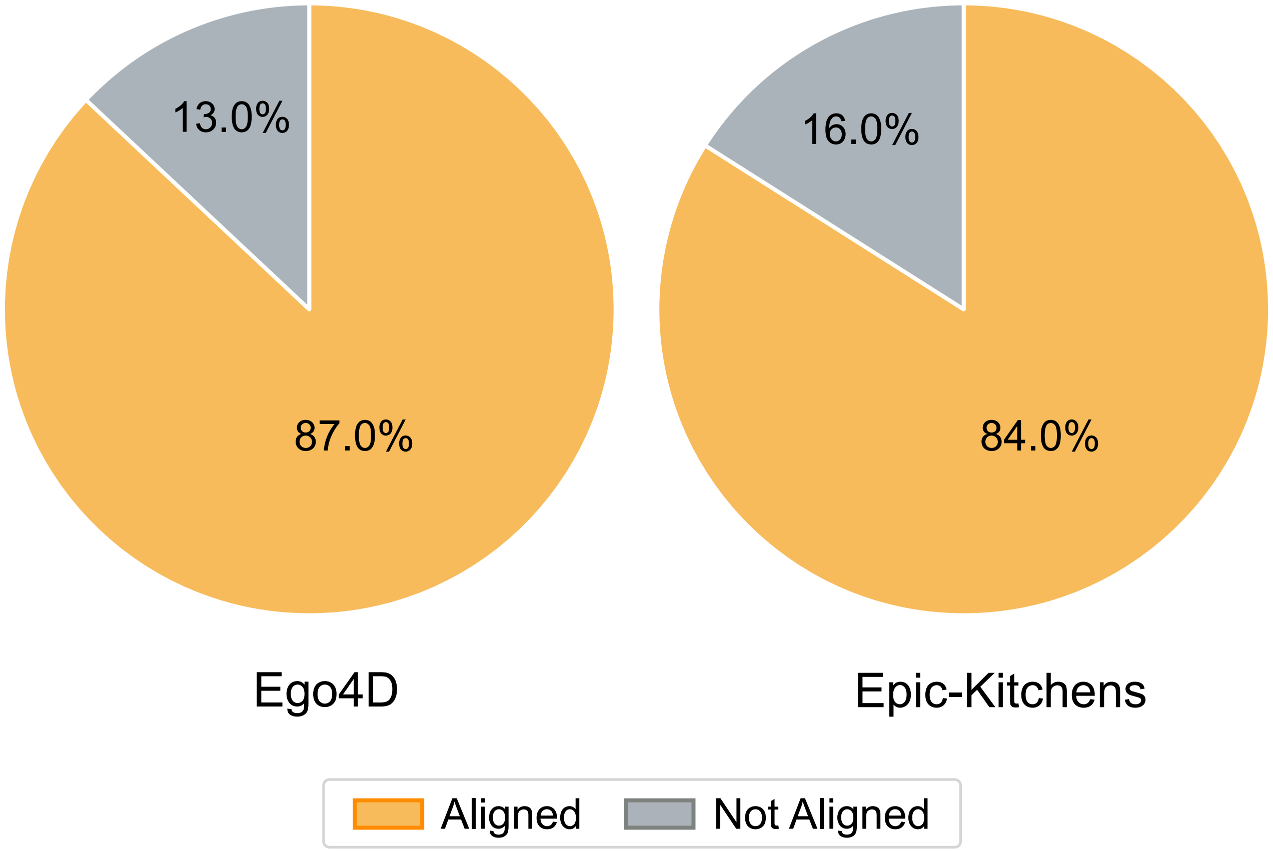
The person uses their left hand to pick up a spoon from the countertop and holds it while standing next to the hob, pan, and sauce.

Select an option

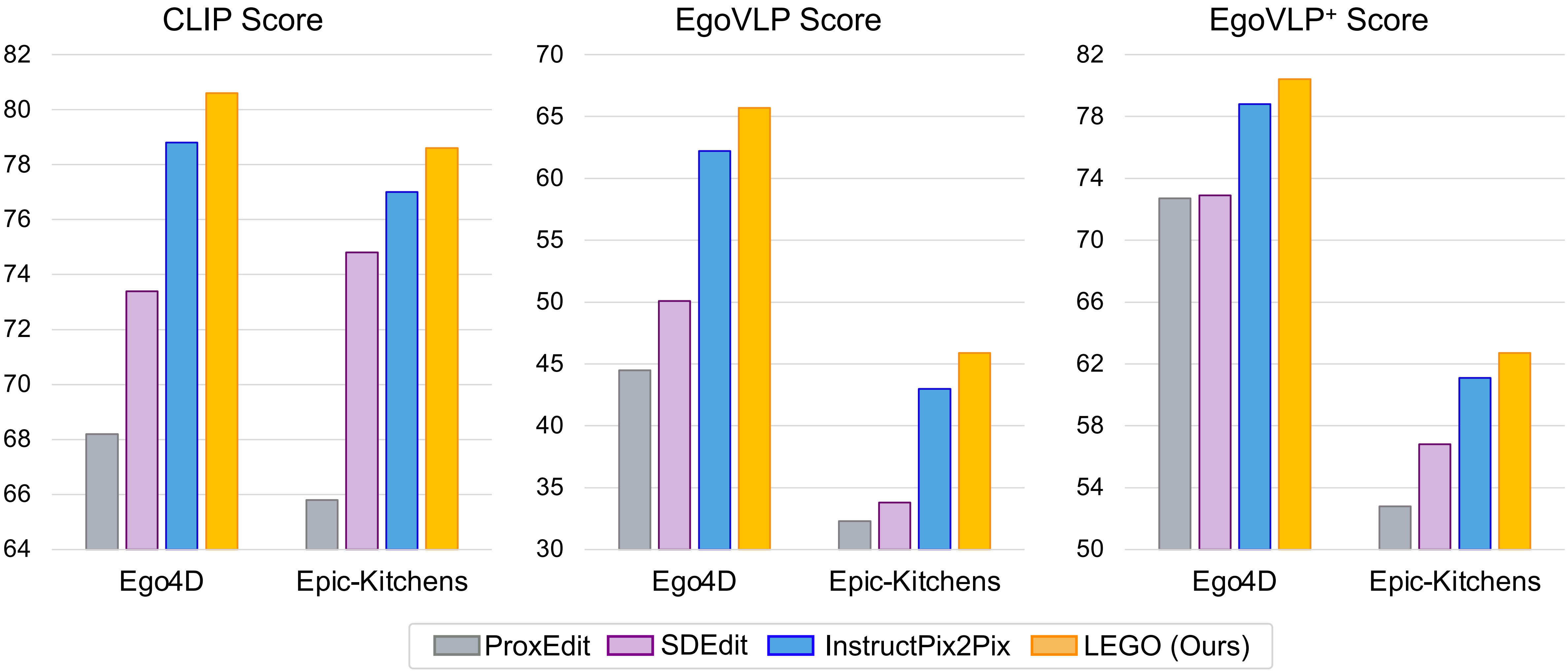
Aligned 1

Not Aligned 2

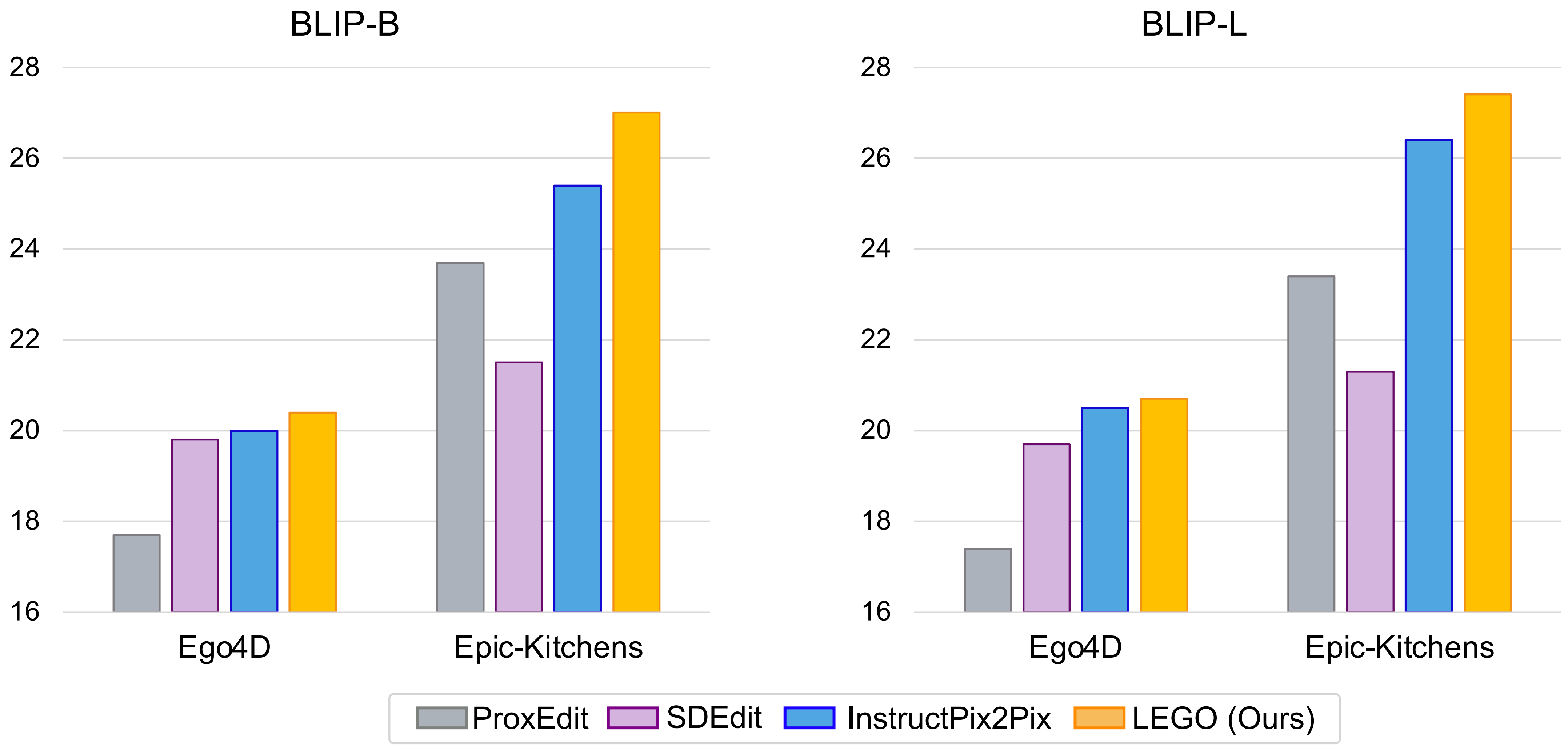
User Study Interface



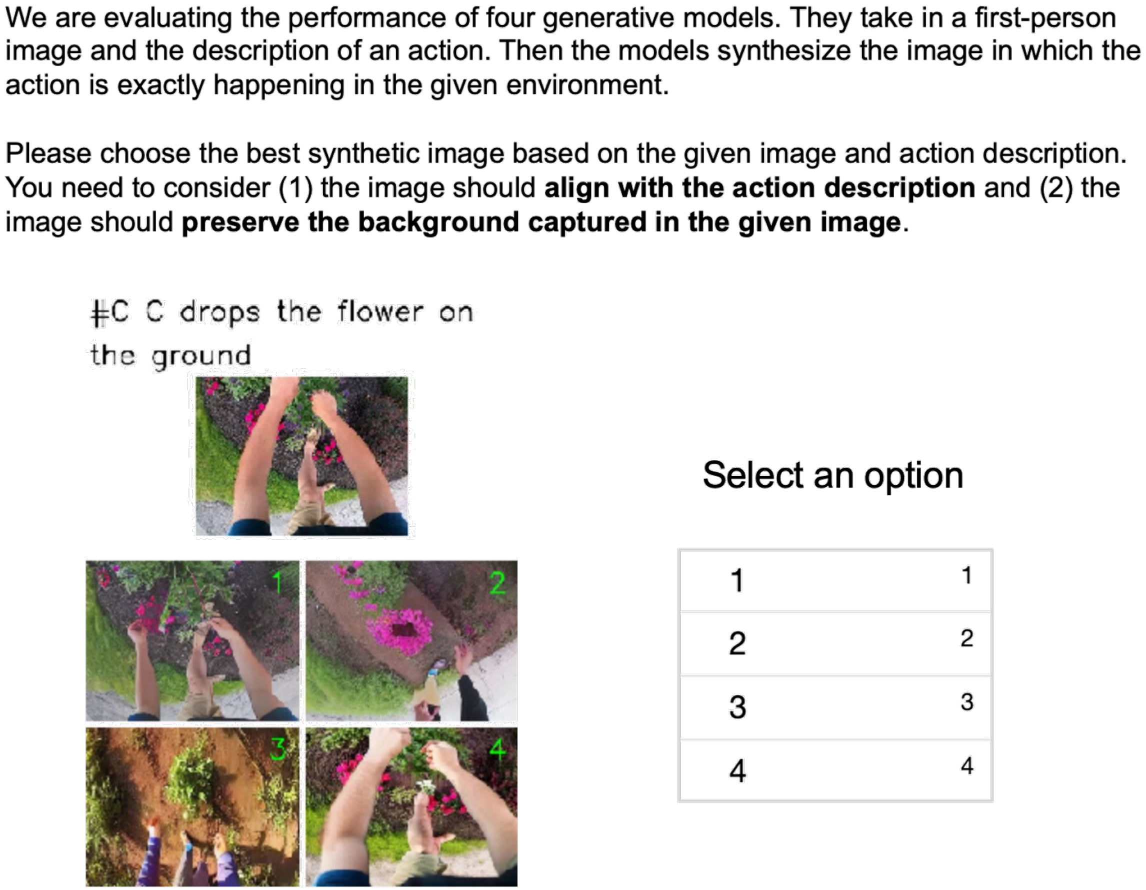
Experiments – Comparison with Prior Methods



Experiments – Comparison with Prior Methods



Experiments – User Study



We are evaluating the performance of four generative models. They take in a first-person image and the description of an action. Then the models synthesize the image in which the action is exactly happening in the given environment.

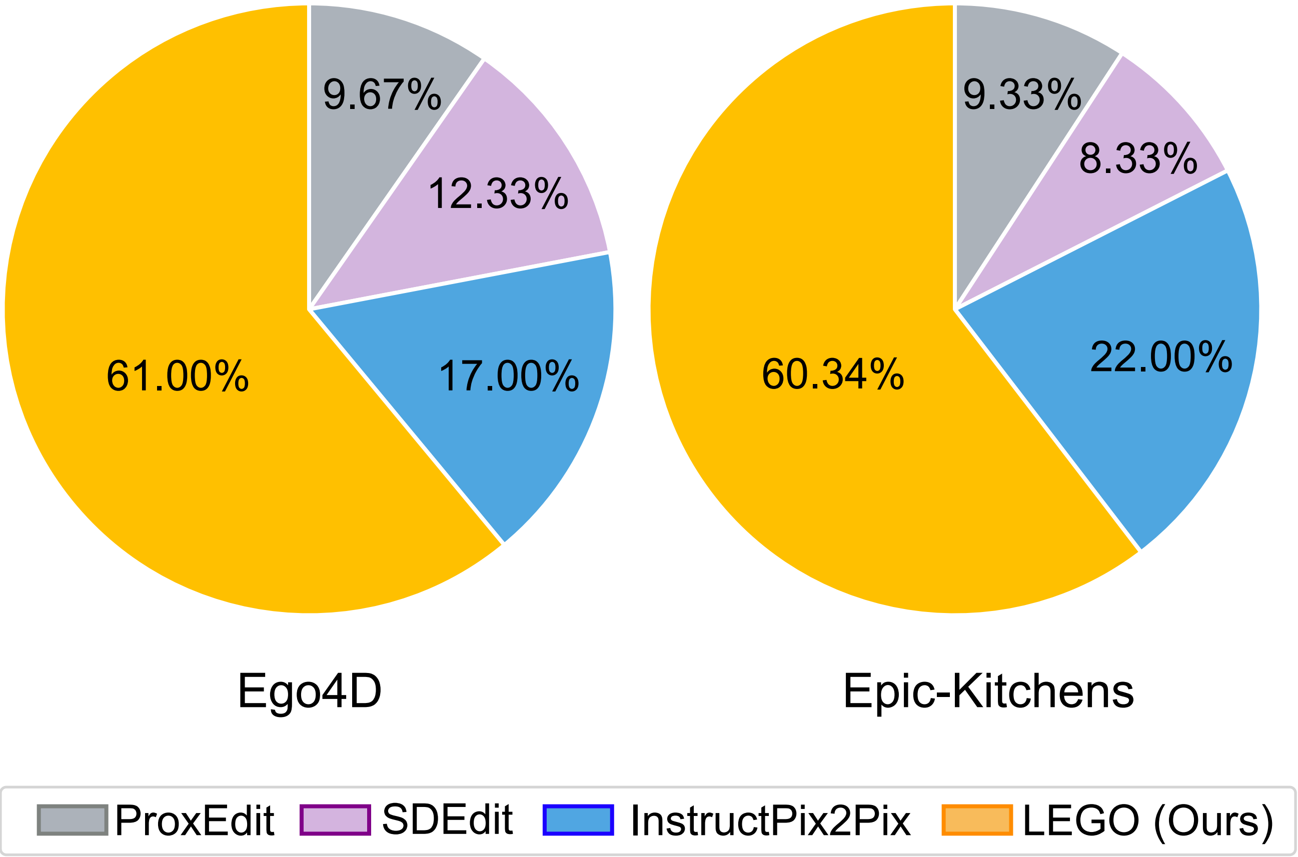
Please choose the best synthetic image based on the given image and action description. You need to consider (1) the image should **align with the action description** and (2) the image should **preserve the background captured in the given image**.

#C C drops the flower on the ground

Select an option

1	1
2	2
3	3
4	4

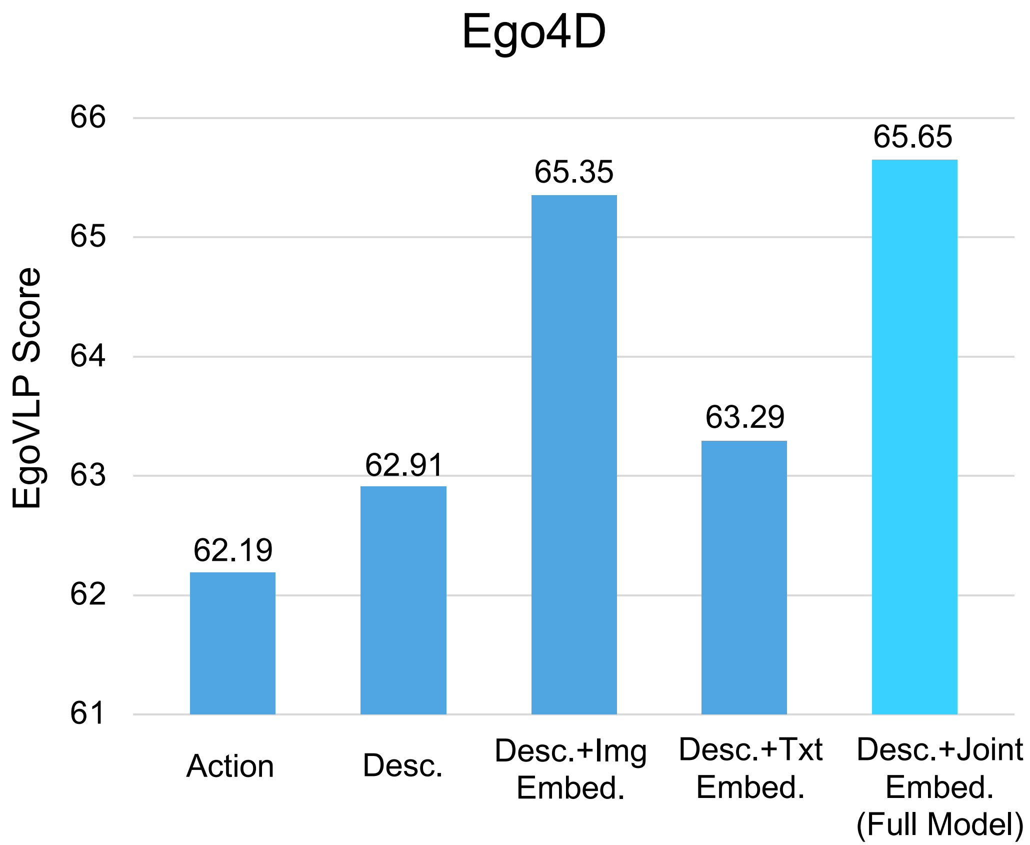
User Study Interface

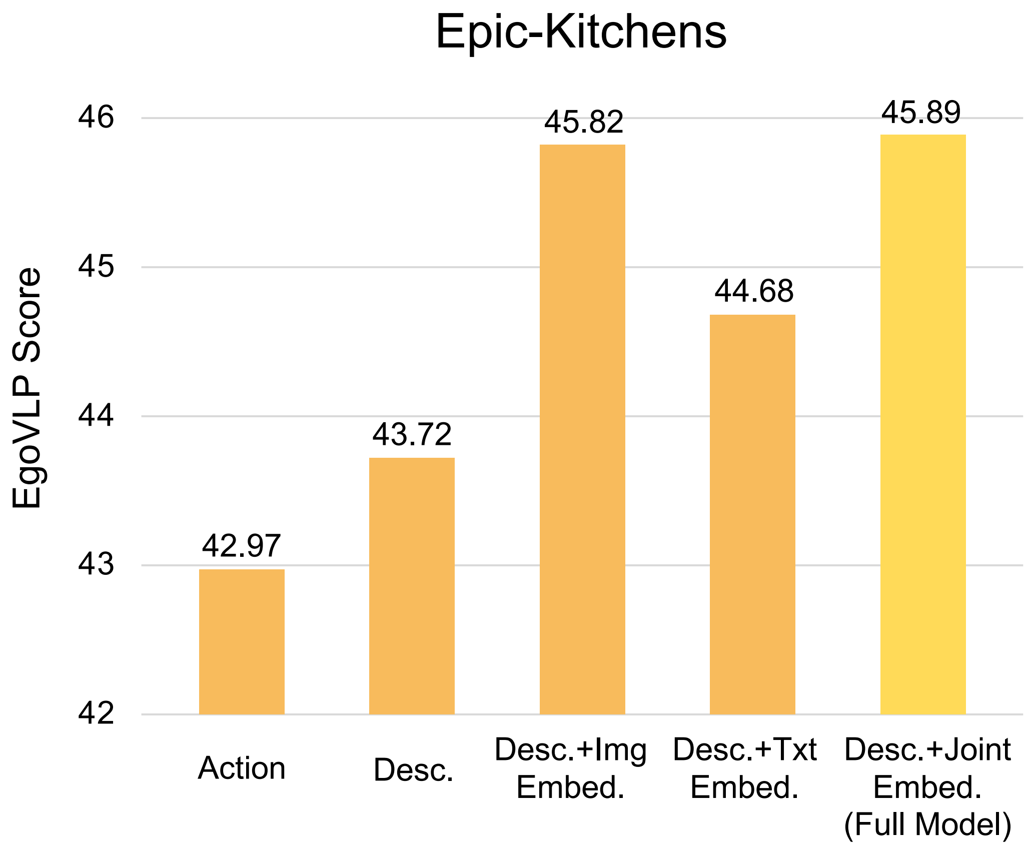


Ego4D

Epic-Kitchens

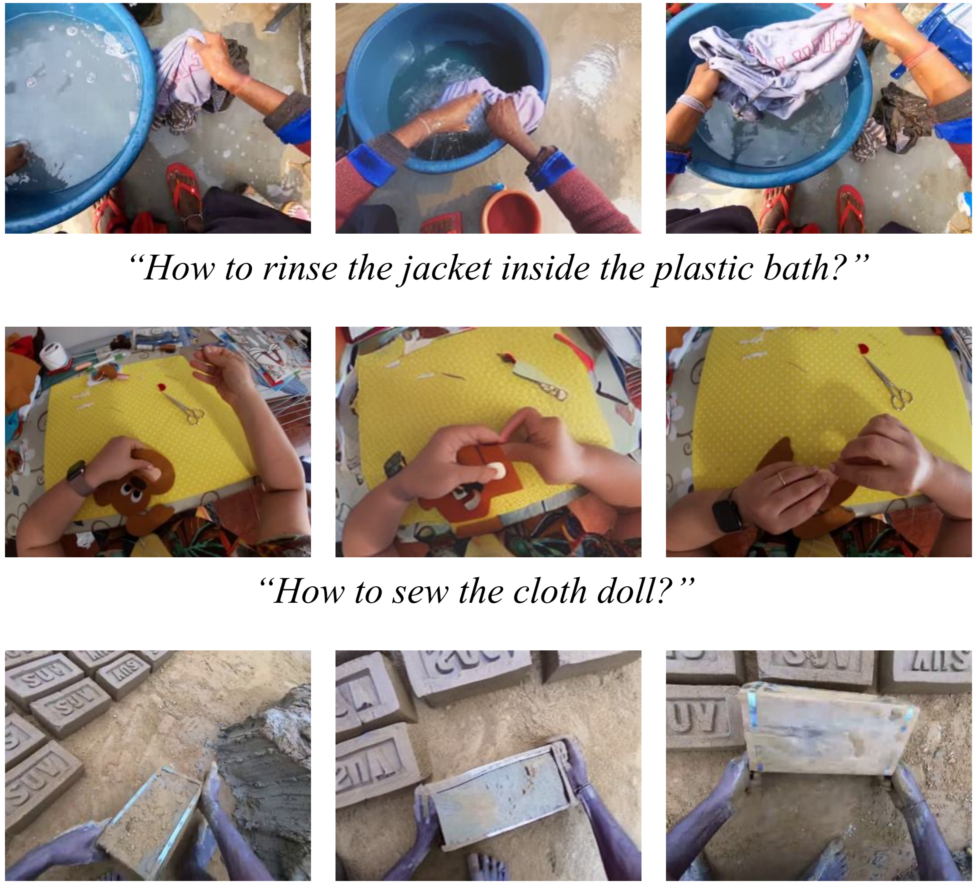
Experiments – Ablation Study


Ego4D


Epic-Kitchens

Experiments – Visualization

Input Frame Generated Frame Ground Truth



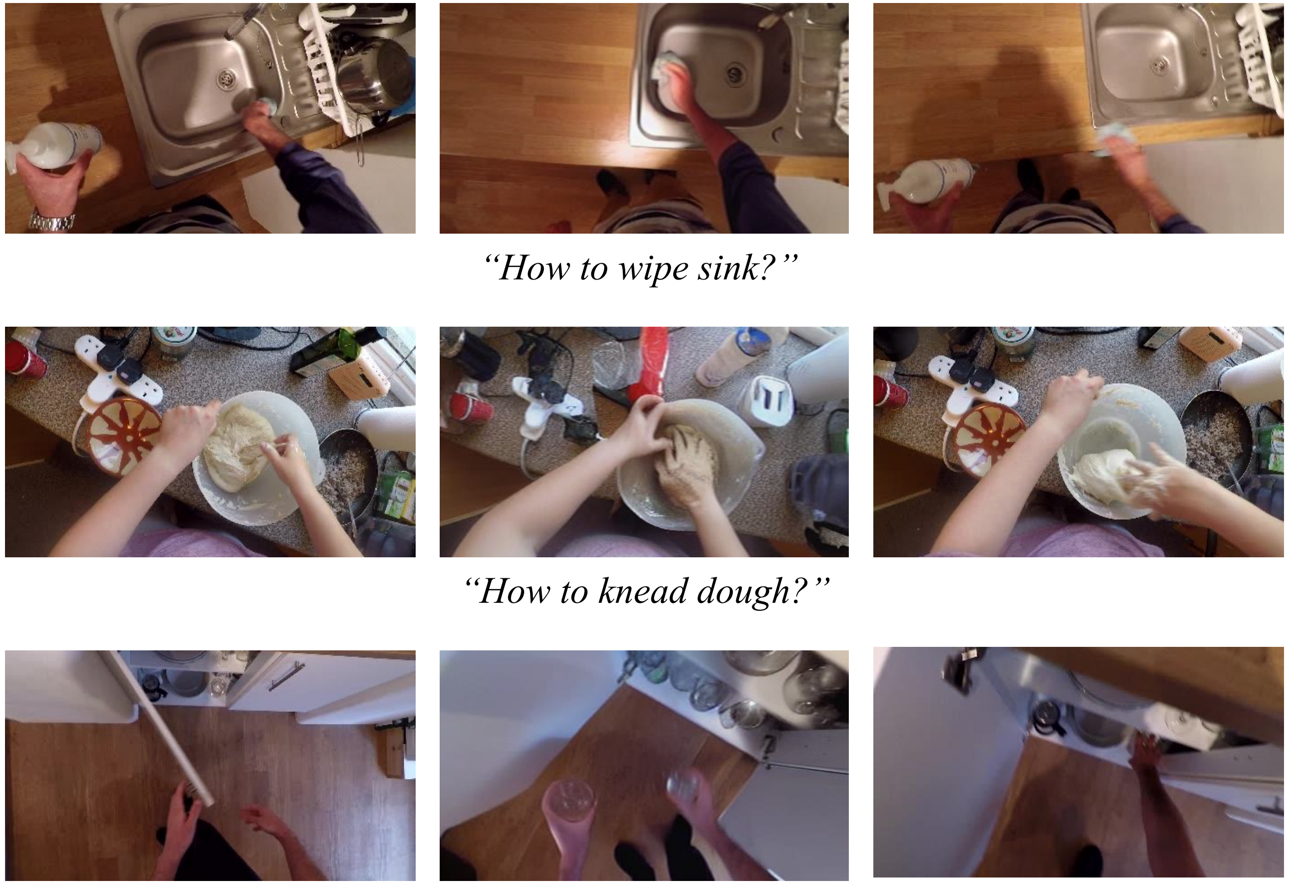
"How to rinse the jacket inside the plastic bath?"

"How to sew the cloth doll?"

"How to pour clay mix from the brick mold on the ground?"

Ego4D

Input Frame Generated Frame Ground Truth



"How to wipe sink?"

"How to knead dough?"

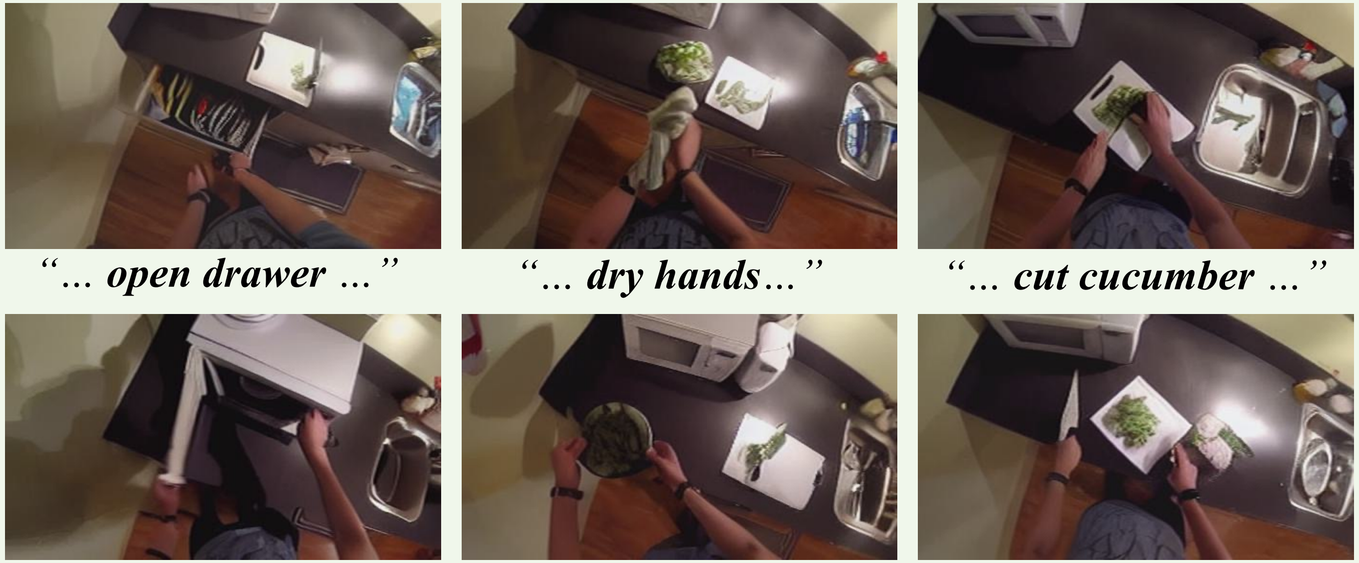
"How to take glass?"

Epic-Kitchens

Experiments – Visualization

Generating diverse actions in the same contexts.

"Can you provide instructions on how to ***{action}*** *in my current situation?"*



"… open drawer …" ***"… dry hands…"*** ***"… cut cucumber …"***

"… open microwave …" ***"… pick up bowl …"*** ***"… take knife …"***

LEGO

Conclusion

- We introduce the novel problem of egocentric action frame generation to facilitate the process of skill transfer.

- We propose a prompt enrichment scheme by visual instruction tuning to help the diffusion model understand the action state transition from the egocentric perspective.

- We propose a novel method to leverage VLLM text and visual embeddings to improve the performance of the latent diffusion model.

- We conduct thorough experiments on Ego4D and Epic-Kitchens datasets to validate the superiority of our method and provide more insights in our model.

Thank You

Our code, data and model weights have been released!

Project Page GitHub Twitter

Our poster will be presented at

# 240

4:30 pm - 6:30 pm

Welcome to stop by our poster
for further discussion!