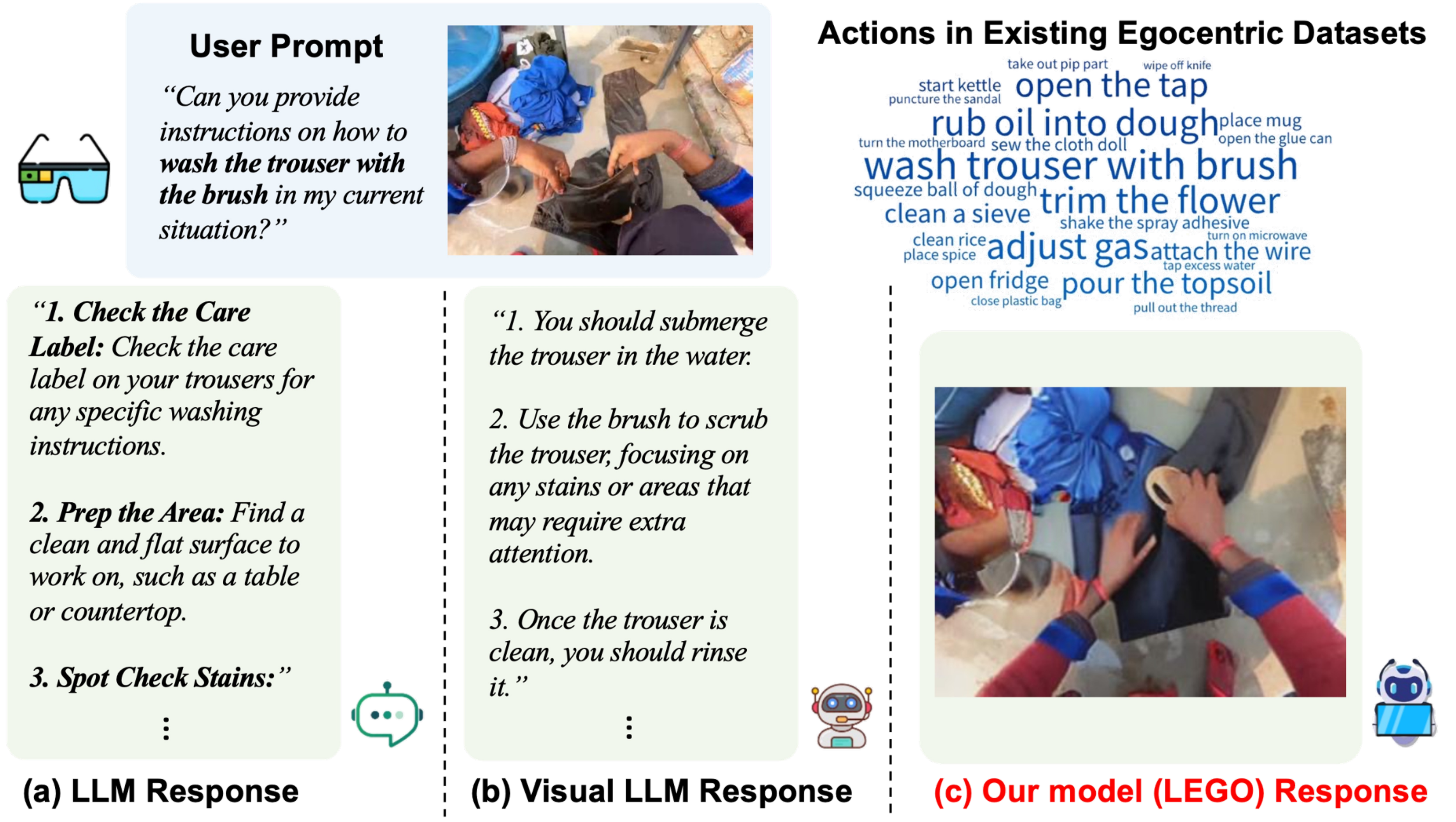
LEGO: Learning EGOcentric Action Frame Generation via Visual Instruction Tuning

Bolin Lai[1,2] Xiaoliang Dai[1] Lawrence Chen[1] Guan Pang[1] James M. Rehg[3] Miao Liu[1]

[1]GenAI, Meta [2]Georgia Tech [3]UIUC

Motivation



User Prompt

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

*"**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:**"*

⋮

"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."

⋮

Actions in Existing Egocentric Datasets

(a) LLM Response **(b) Visual LLM Response** **(c) Our model (LEGO) Response**

When a user asks for instructions on a task:

- *LLM* -- the answer is too generic and verbose, which is hard to follow.
- *Visual LLM* -- she still faces the challenge of parsing a written description.
- *LEGO (our method)* -- generates an image that provides visual guidance exactly in her situation from the egocentric viewpoint.

We thus propose a new task -- Egocentric Action Frame Generation,

Input: (1) User query of how to perform an action, (2) An image of current situation before an action happens. **Output:** An image in which the action is being performed.

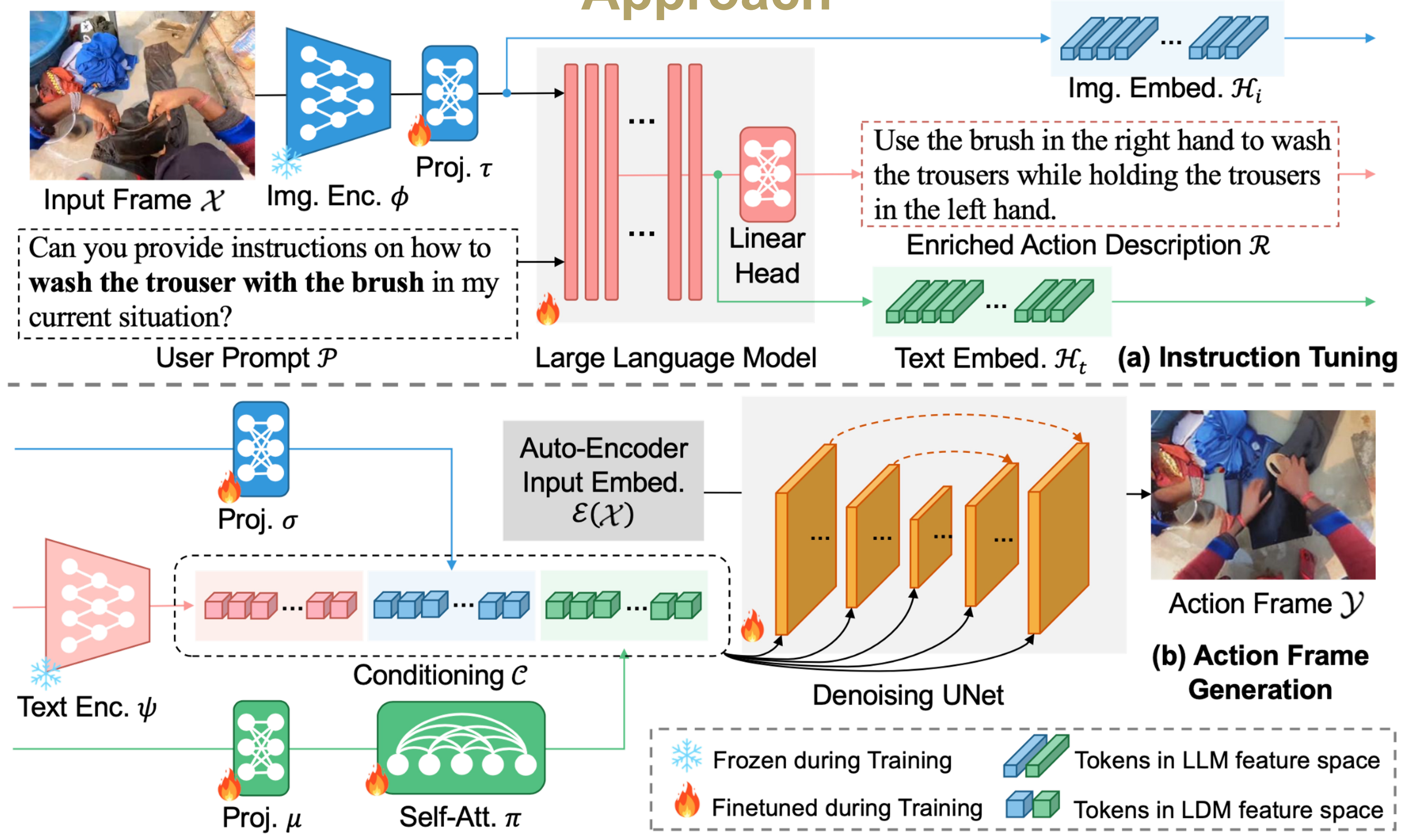
Challenges

- 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.

→

- Enriching the action labels with LLM via visual instruction tuning.
- Leveraging finetuned LLM embeddings to improve egocentric action frame generation.

Approach



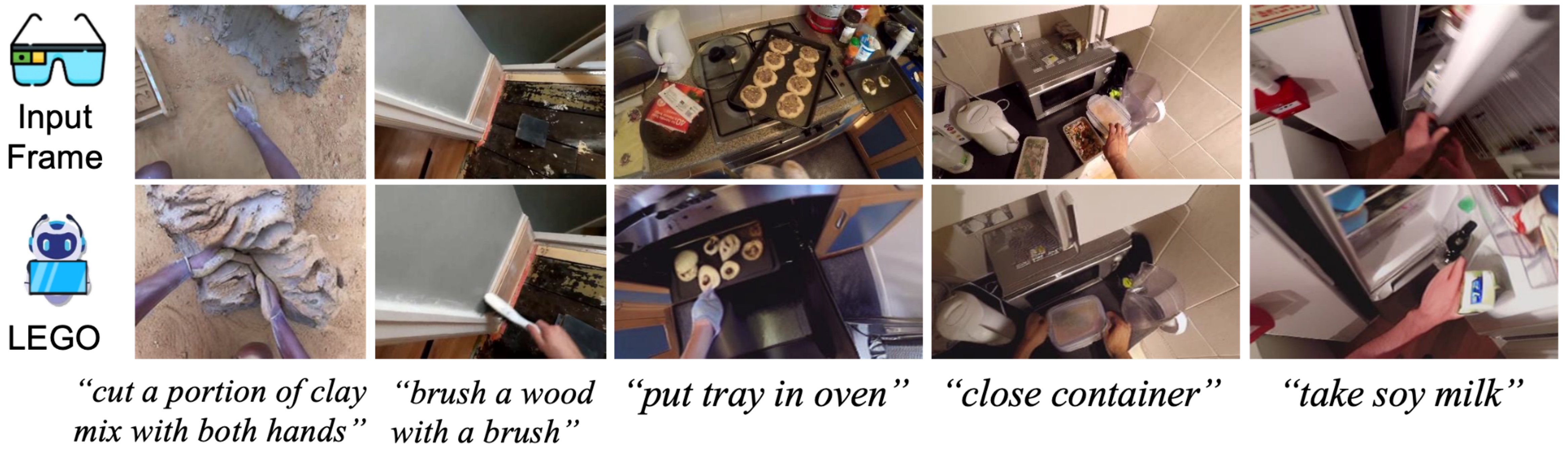
Input Frame $\mathcal{X}$
Img. Enc. ϕ
Proj. τ
Large Language Model
Linear Head
Img. Embed. $\mathcal{H}_i$
Enriched Action Description $\mathcal{R}$
Text Embed. $\mathcal{H}_t$
User Prompt $\mathcal{P}$
(a) Instruction Tuning
Proj. σ
Auto-Encoder Input Embed. $\mathcal{E}(\mathcal{X})$
Text Enc. ψ
Conditioning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

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.

LEGO consists of two key components:

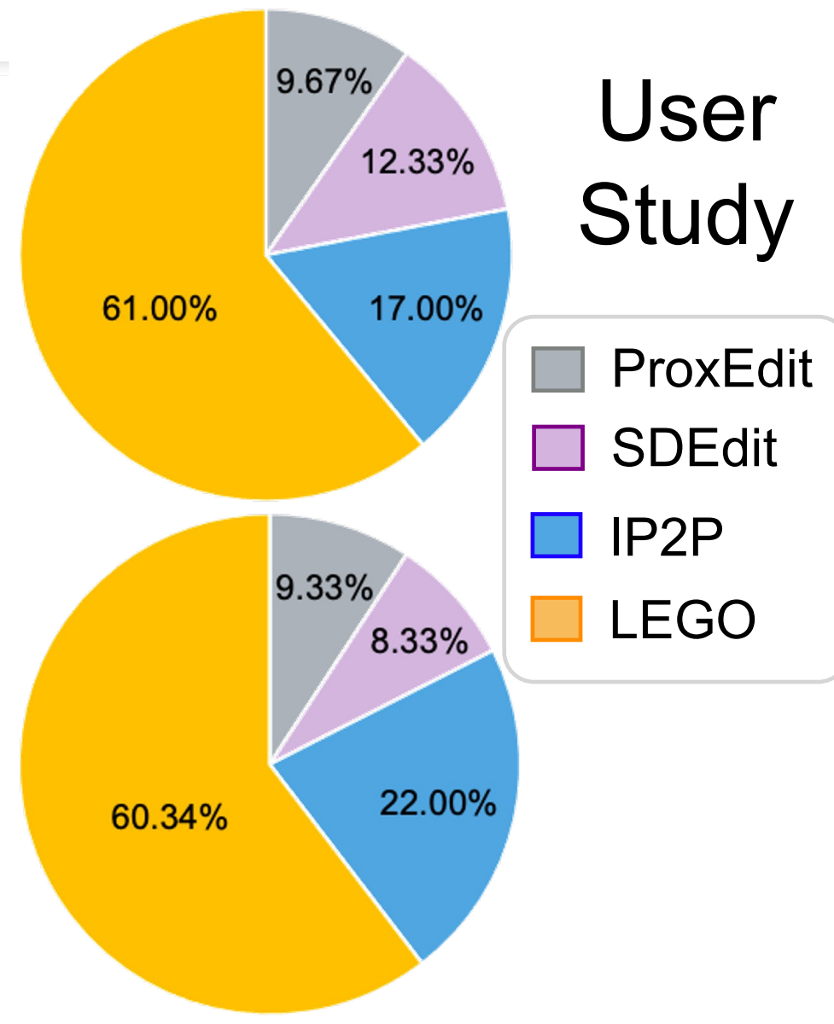
- *Visual Instruction Tuning* -- We finetune an LLM to generate detailed action descriptions which include information such as hands and spatial locations.
- *Action Frame Generation* -- We project image and text features from LLM to LDM space, and input them to a diffusion model as additional conditions to mitigate the domain gap.

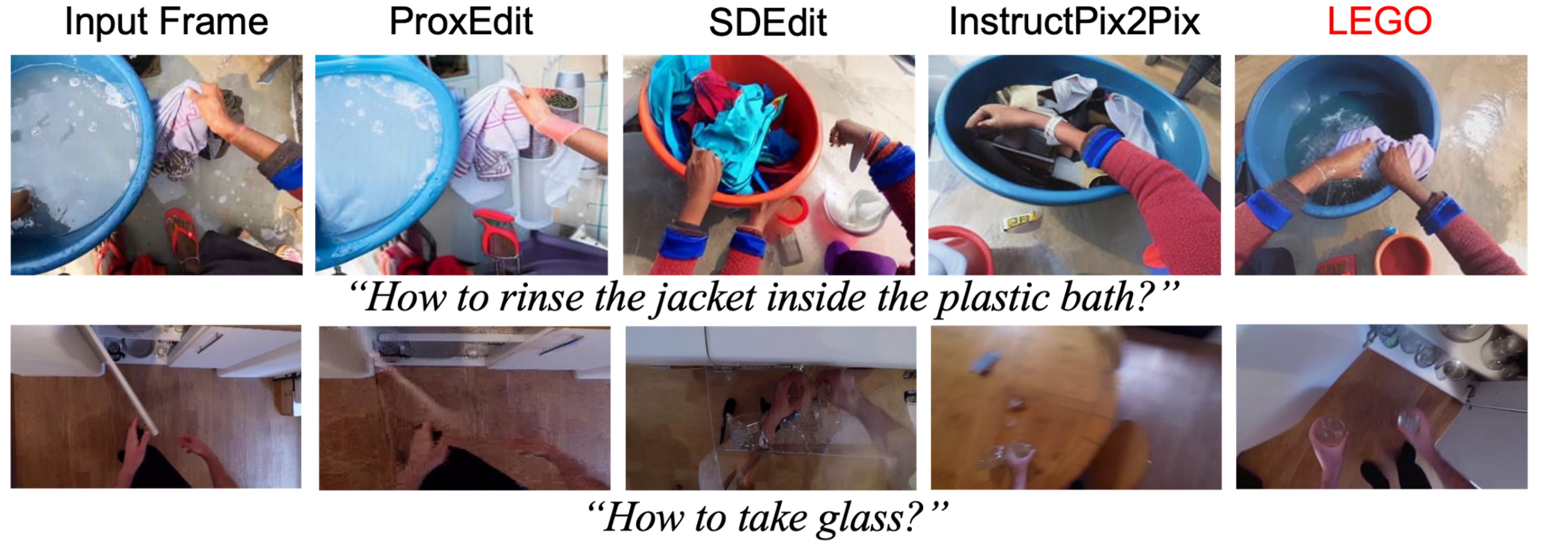


"cut a portion of clay mix with both hands" *"brush a wood with a brush"* *"put tray in oven"* *"close container"* *"take soy milk"*

Experiments and Results

	Methods	EgoVLP	EgoVLP^{+}	CLIP	FID ↓	PSNR	LPIPS ↓
Ego4D	ProxEdit [26]	44.51	72.68	68.17	33.01	11.88	40.90
	SDEdit [59]	50.07	72.90	73.35	33.35	11.81	41.60
	IP2P [6]	62.19	78.84	78.75	24.73	12.16	37.16
	LEGO	**65.65**	**80.44**	**80.61**	**23.83**	**12.29**	**36.43**
E-Kitchens	ProxEdit [26]	32.27	52.77	65.80	51.35	11.06	46.35
	SDEdit [59]	33.84	56.80	74.76	27.41	11.30	43.33
	IP2P [6]	42.97	61.06	77.03	**20.64**	11.23	40.82
	LEGO	**45.89**	**62.66**	**78.63**	21.57	**11.33**	**40.36**

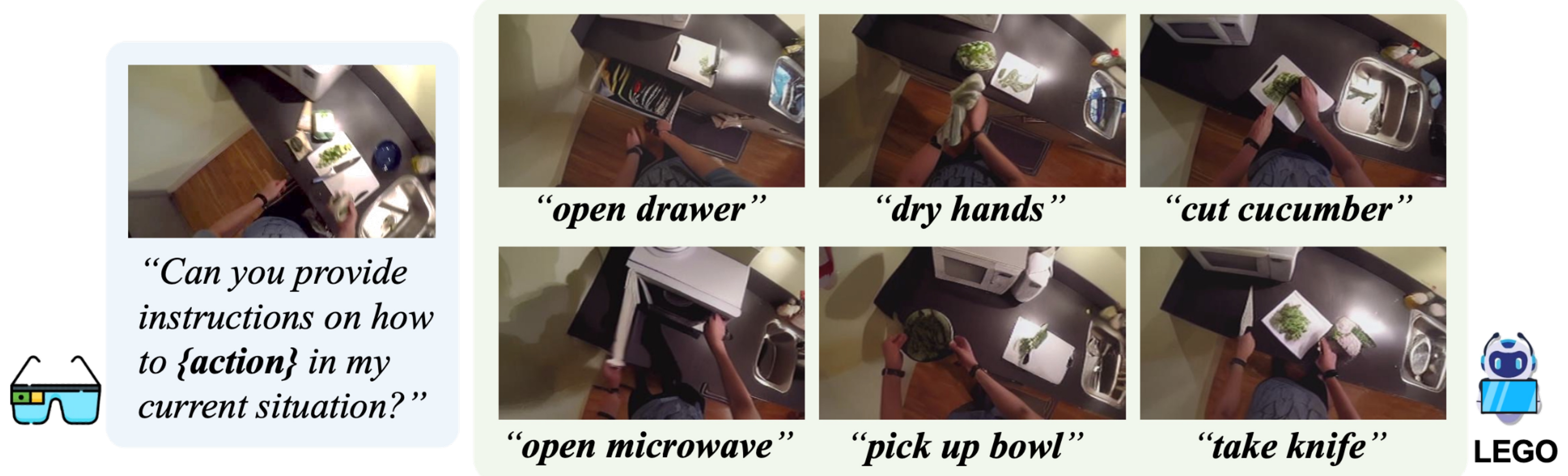




"How to rinse the jacket inside the plastic bath?"

"How to take glass?"

Generating various actions in the same contexts:



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

Contact

Homepage WeChat Twitter Codes