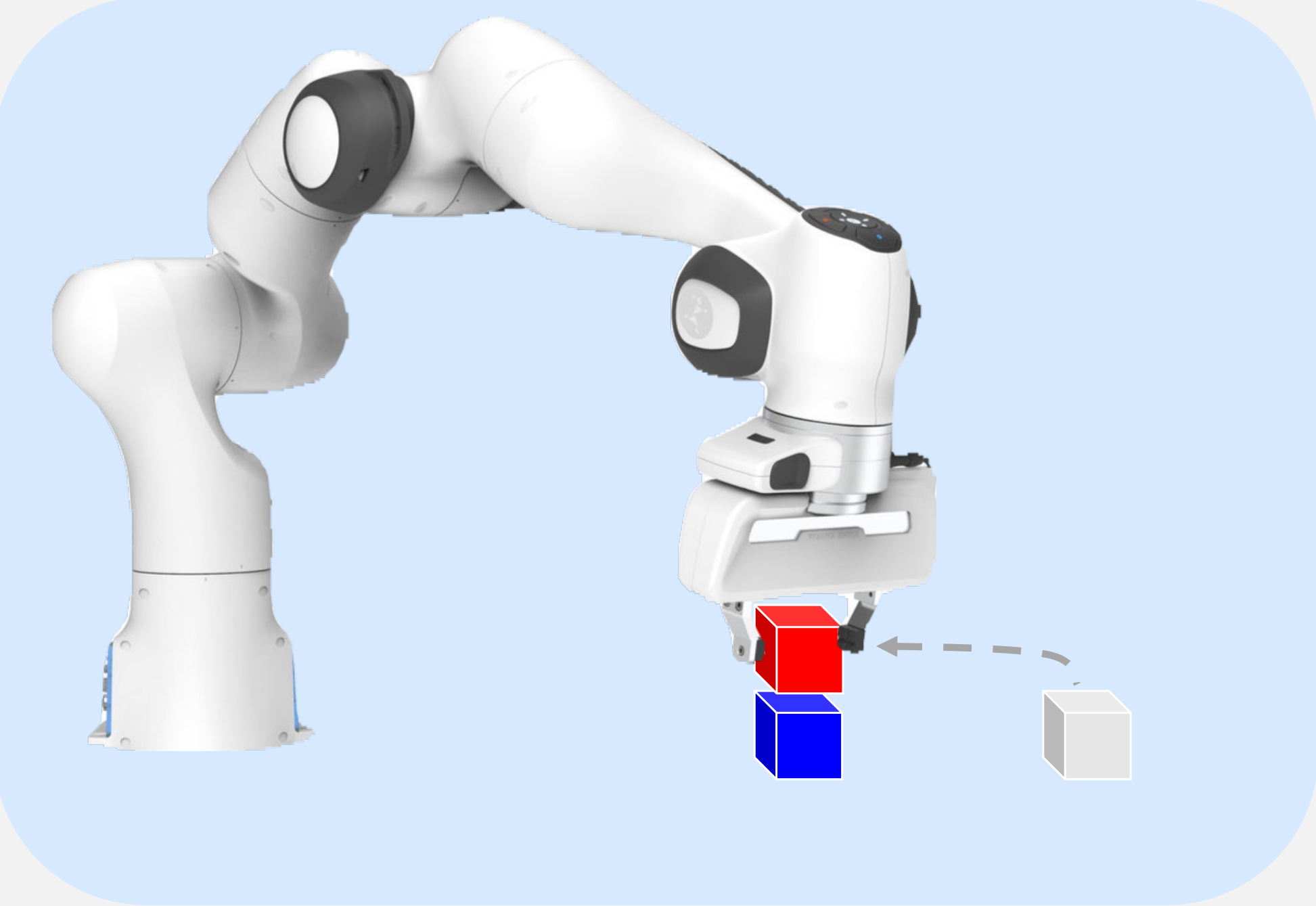
Collaborative Robotics Lab

GLOMA: Grounded Location for Object Manipulation

Yifan (Brandon) Yang, Mohammad Yasar, Tariq Iqbal

University of Virginia

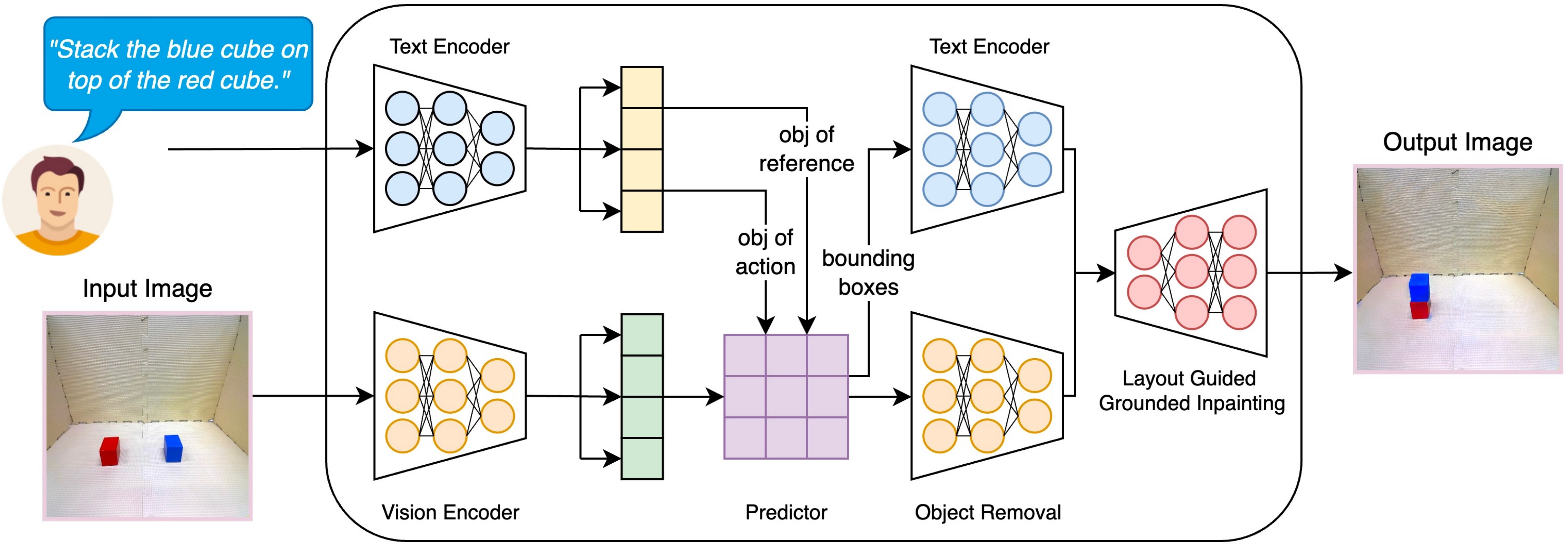
Motivation



- Solving Robotics Tasks with goal-conditioned RL.
- Require goal state represented as an image.

- Existing diffusion models proficient at style transfers, text-to-image prompts.
- Current models unable to achieve real-time object location manipulation.

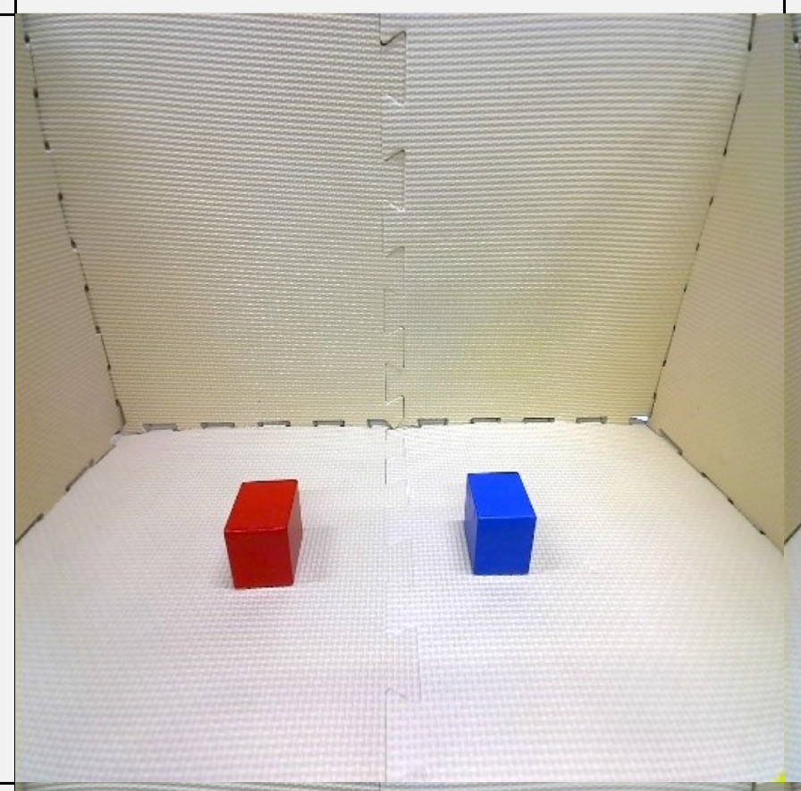
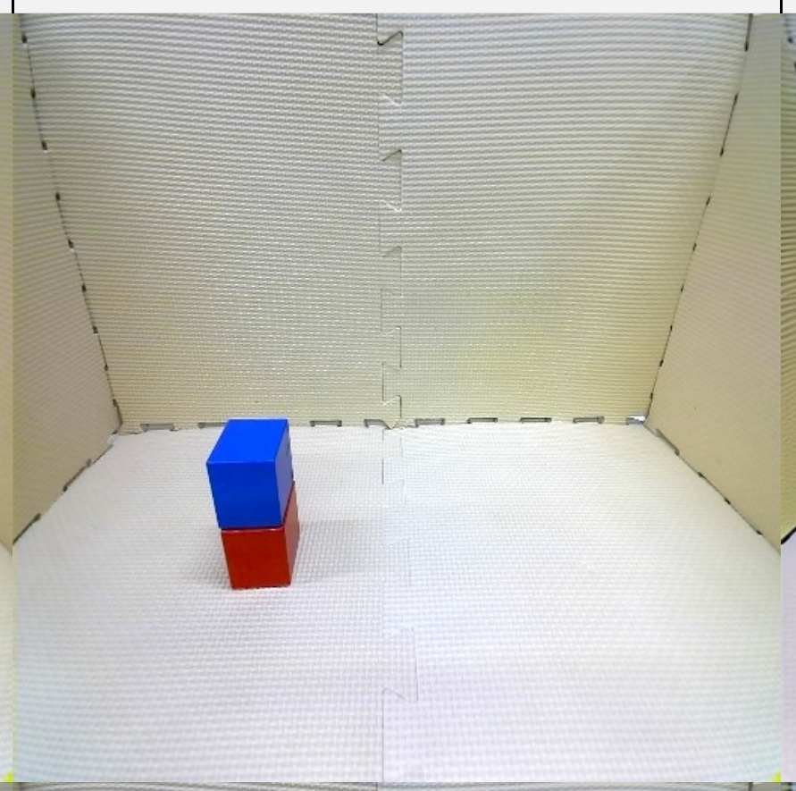
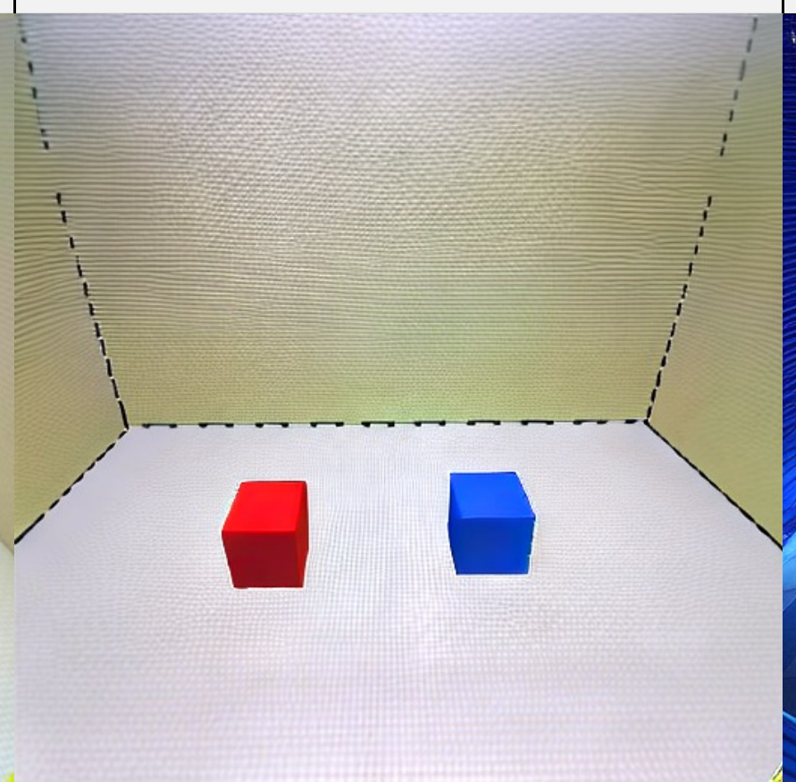
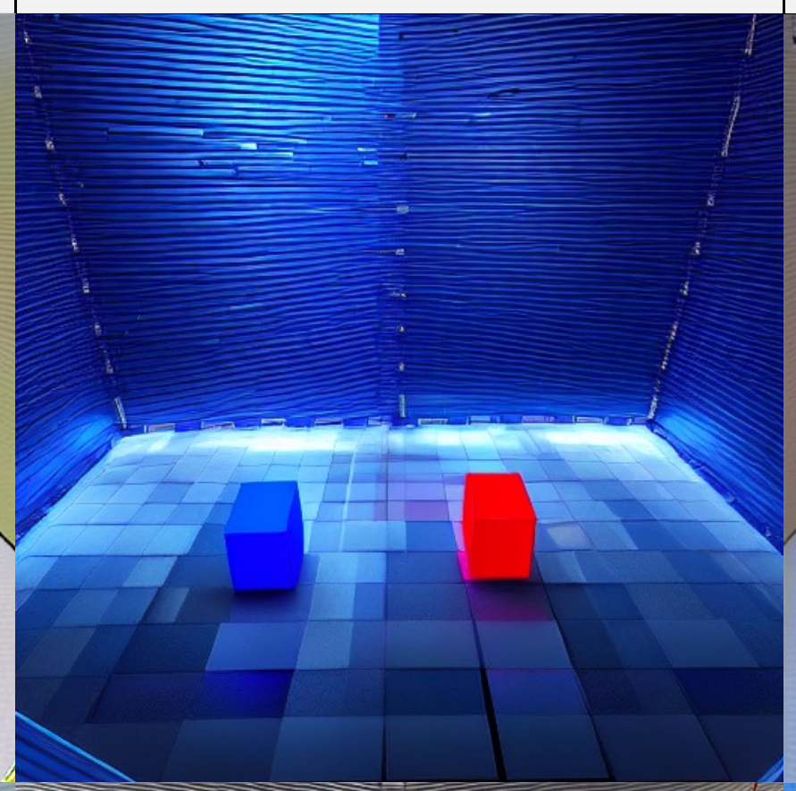
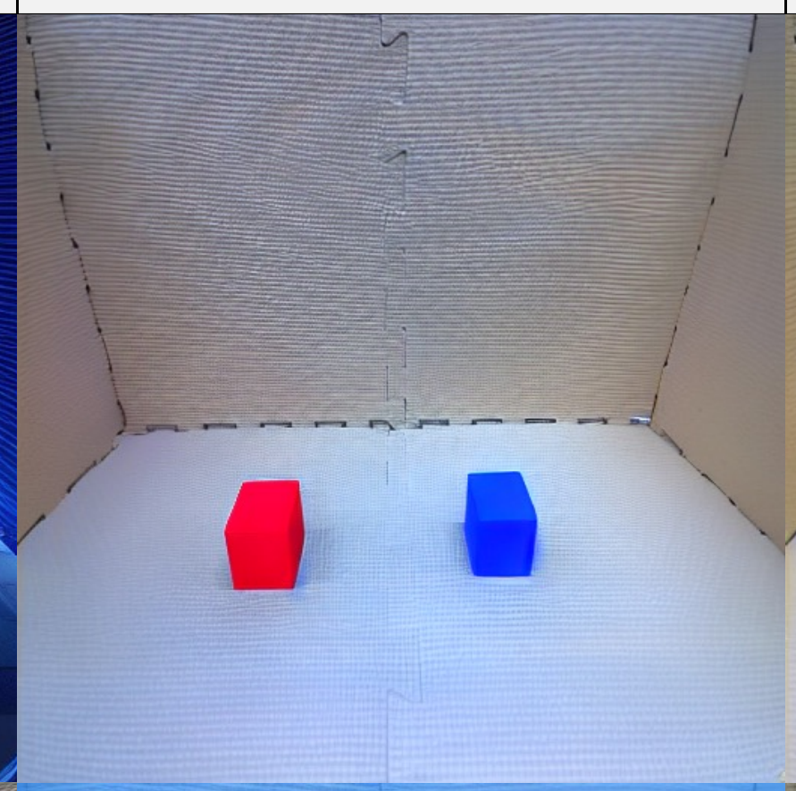
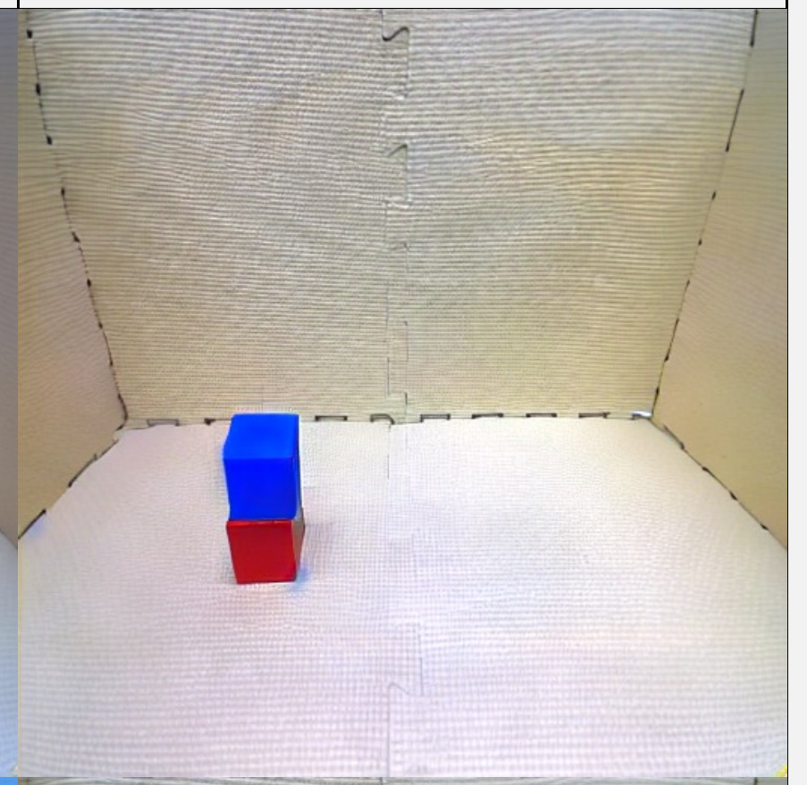
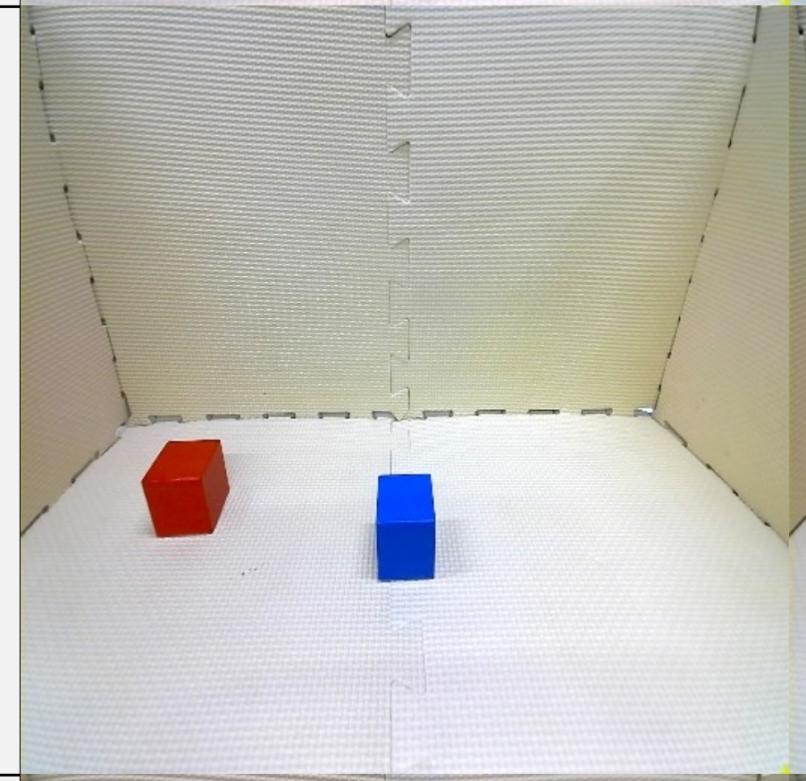
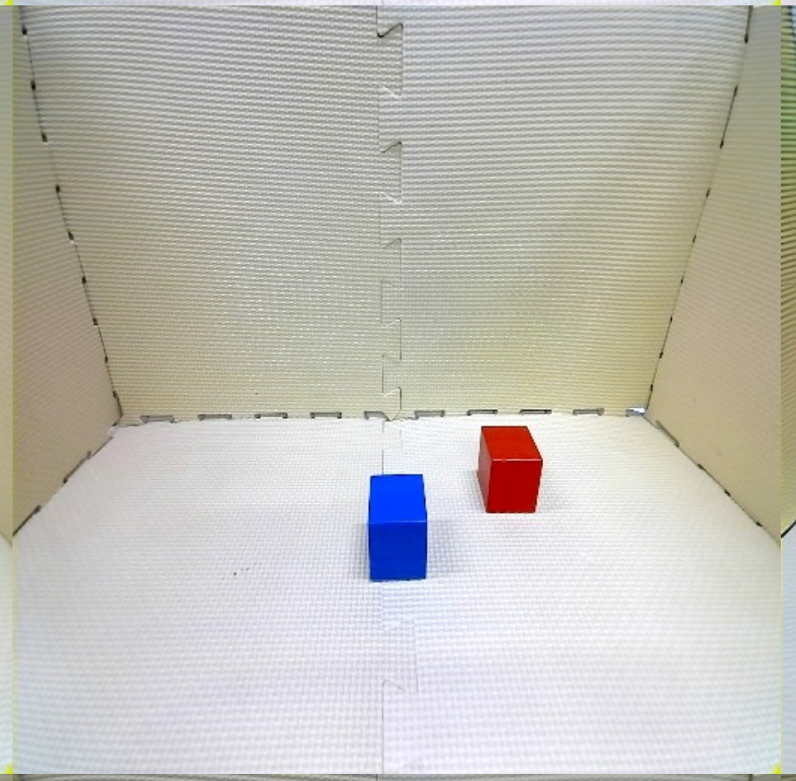
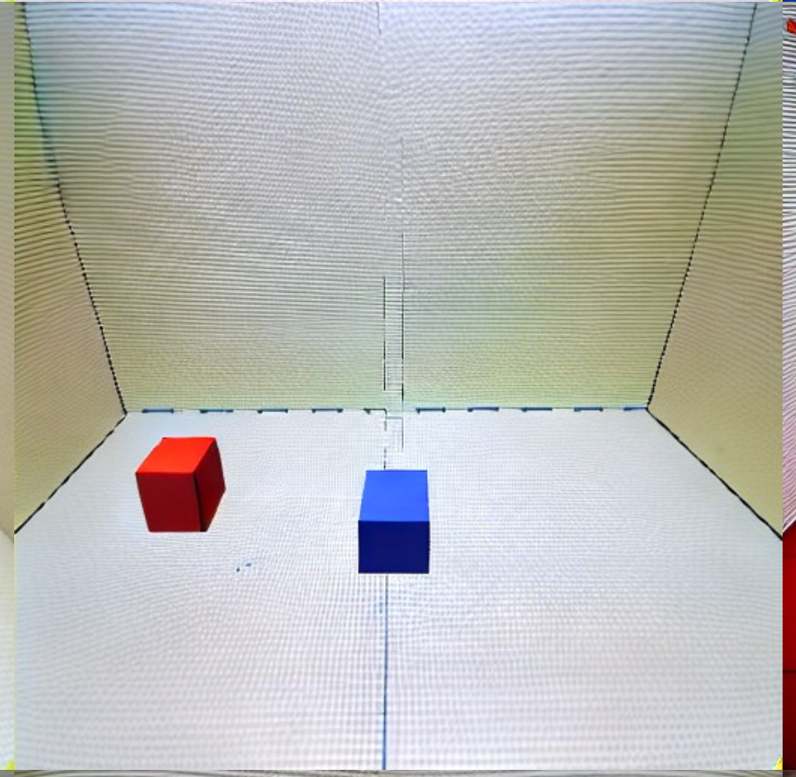
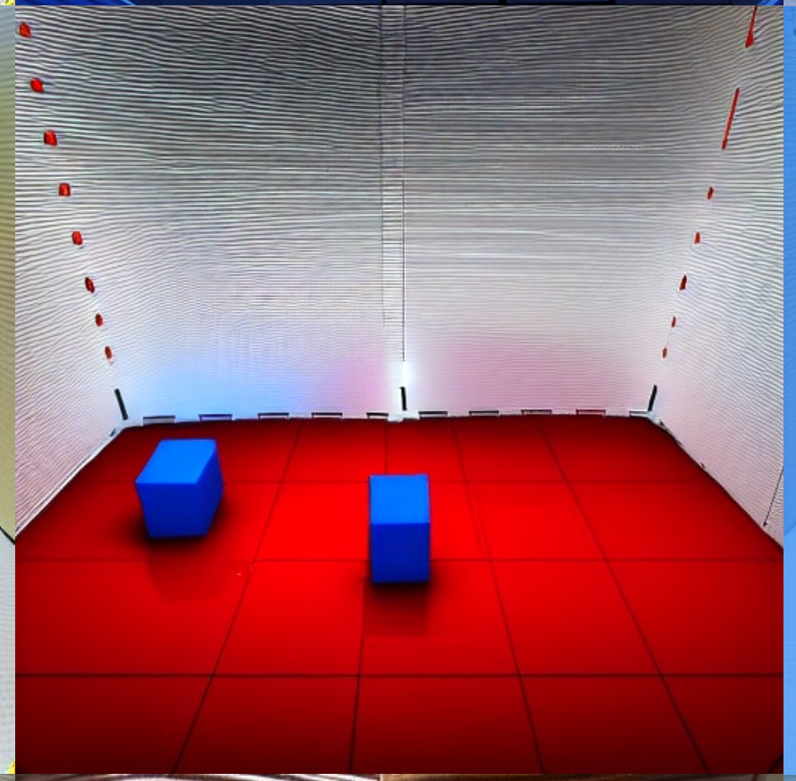
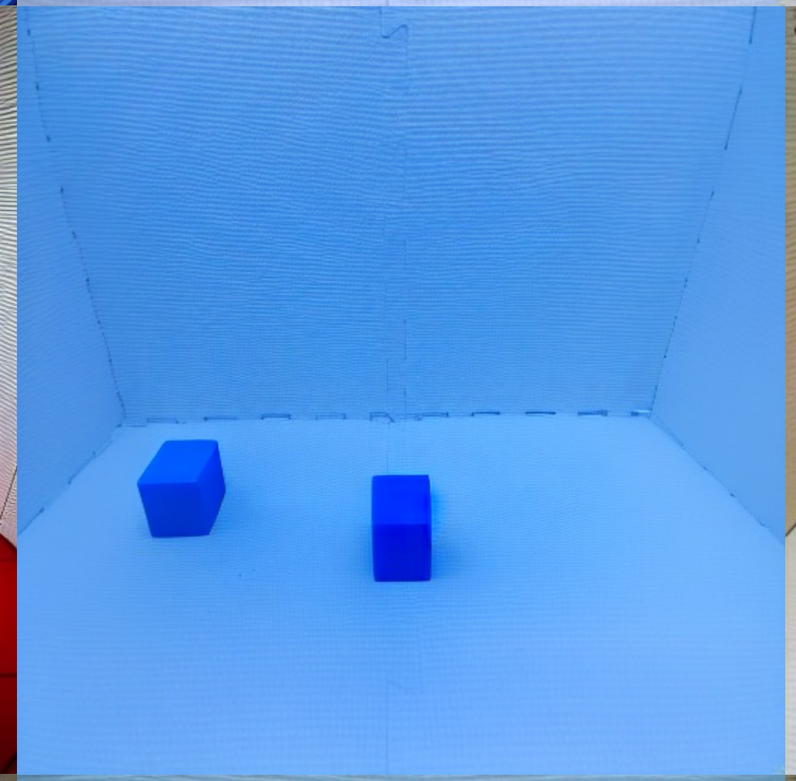
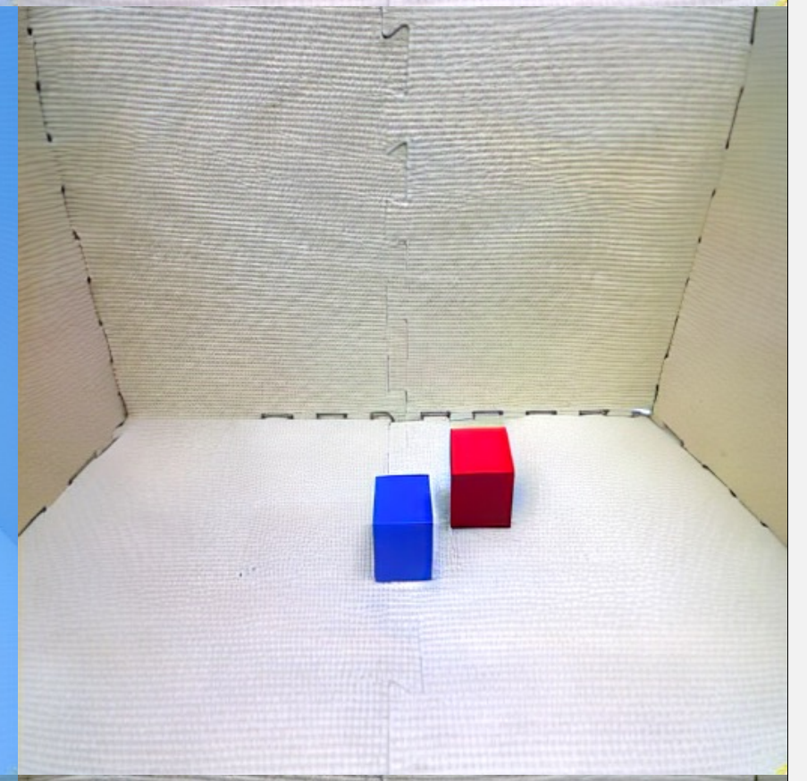
Architecture



Quantitative Results

Model	Task	SSIM ↑	VGG16 feature similarity ↑	FID Score ↓
LEDITS	Stacking	0.74	0.48	299.085
	Moving	0.76	0.59	
	Other Objects	0.73	0.26	
ControlNet	Stacking	0.29	0.27	426.548
	Moving	0.31	0.29	
	Other Objects	0.37	0.22	
InstructPix2Pix	Stacking	0.84	0.60	308.686
	Moving	0.80	0.49	
	Other Objects	0.76	0.13	
GLOMA (Ours)	Stacking	**0.86**	**0.78**	**190.119**
	Moving	**0.86**	**0.75**	
	Other Objects	**0.83**	**0.48**	

Qualitative Results

Edit Instruction	Input Image	**Ground Truth**	LEDITS	ControlNet	InstructPix2Pix	**GLOMA (Ours)**
Move the blue cube on top of the red cube.						
Move the red cube to the right of the blue cube.						
Put the football on top of the cup.