A. Examples of subgoal generation with different sizes of training set

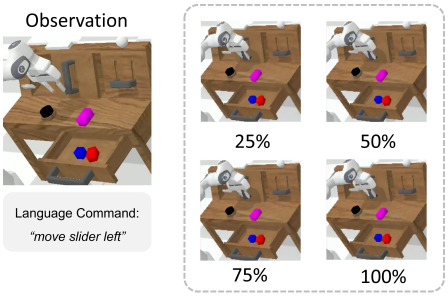

Figure 7. The generated subgoals using models trained with different amounts of data. The generated subgoals are nearly identical across varying dataset sizes.

B. Example of subgoal generation on unseen tasks and environment

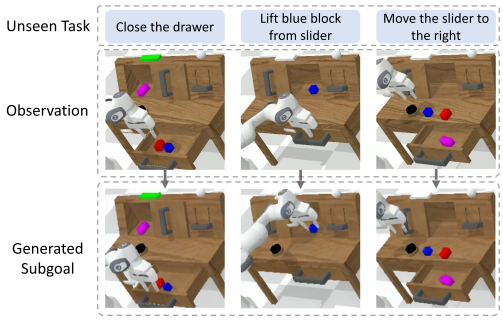

Figure 8. Example rollouts demonstrate how the generated subgoals guide the robot to perform unseen tasks.

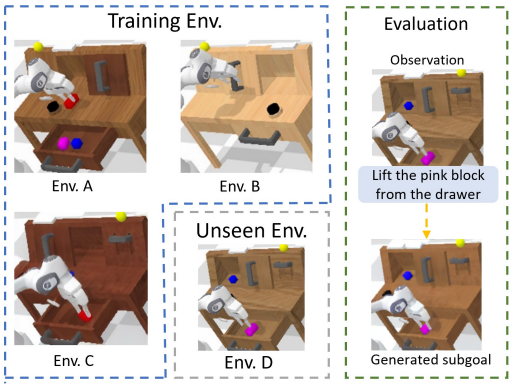

Figure 9. TaKSIE is trained using environment A, B, and C, and then tested on an unseen environment D. The example on the right demonstrates that even in an unseen environment, the model is capable of generating realistic subgoals.

C. Example of failures cases

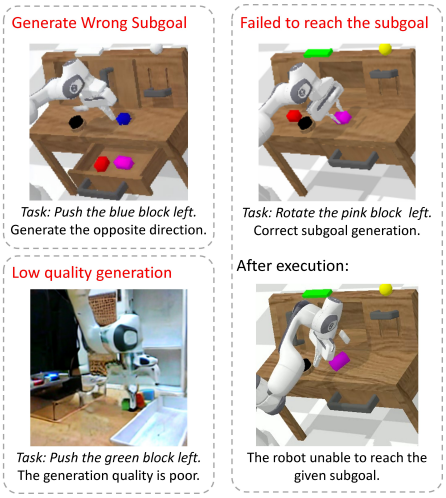

Figure 10. Three common failure cases: (1) incorrect subgoals, (2) low-quality images, and (3) failure to reach valid subgoals.

D. Comparsion of different image-conditioned diffusion model

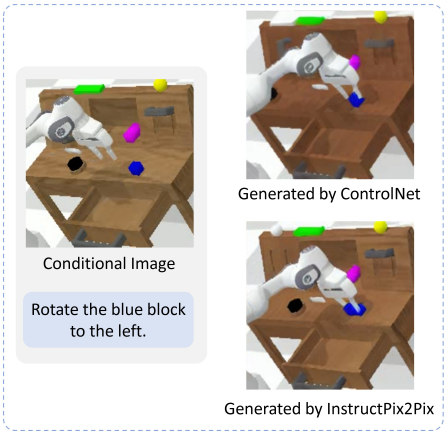

Figure 11. Comparison of subgoal generation using different image-conditioned diffusion models.

In our tests, both InstructPix2Pix and ControlNet demonstrate similar performance in seen environments. However, for unseen environments, we observe that InstructPix2Pix tends to perform better at following the conditioned environment. Fig. 11 provides an example comparison: both ControlNet and InstructPix2Pix are trained on environments A, B, and C from CALVIN and tested on the unseen environment D. InstructPix2Pix generates subgoals that appear more aligned with the conditions of environment D, while ControlNet is more aligned with environ-

δ_1	δ_2	Avg. Num. of Selected Subgoals	Avg. SR (%)
0.02	-0.02	1.02	85.51
0.01	-0.01	1.07	83.99
0.002	-0.002	1.61	76.17
0.001	-0.001	2.03	86.57

Table 6. Number of selected subgoals and success rate (SR) of different slope values for δ_1 and δ_2. The smaller δ_1 and larger δ_2 lead to select more subgoals.

ment C, showing less effective generalization to unseen environments. This observation suggests that InstructPix2Pix may have an advantage in generalizing to new scenarios.

E. Ablation: Slope Values

The slope parameters can affect the selection of subgoals. We experiment with different δ_1 and δ_2 values in all tasks of the CALVIN validation set. Tab. 6 shows that smaller δ_1 and larger δ_2 lead to capture more subgoals. Overall, the number of our selected subgoals is less than SuSIE which selects 3 subgoals on average. Compared to SuSIE's success rate (79.73% in Tab. 4), most of our slope values can perform better, indicating that our selected subgoals can still guide the policy. However, our experiments show no direct relationship between the number of selected subgoals and the success rate as the performance depends on the informativeness of the selected subgoals rather than the number of subgoals.