

References

- [1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. S. Ryoo, G. Salazar, P. R. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. H. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi:10.15607/RSS.2023.XIX.025.
- [2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.
- [3] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.
- [4] C. Lynch and P. Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.
- [5] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.
- [6] Y. Cui, S. Karamcheti, R. Palleti, N. Shivakumar, P. Liang, and D. Sadigh. No, to the right: Online language corrections for robotic manipulation via shared autonomy. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 93–101, 2023.
- [7] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022.
- [8] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022.
- [9] M. Reuss, M. Li, X. Jia, and R. Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *Robotics: Science and Systems (RSS)*, 2023.
- [10] Y. Ding, C. Florensa, P. Abbeel, and M. Phielipp. Goal-conditioned imitation learning. *Advances in neural information processing systems*, 32, 2019.
- [11] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.
- [12] L. Manuelli, W. Gao, P. Florence, and R. Tedrake. kpm: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.
- [13] K. Zakka, A. Zeng, J. Lee, and S. Song. Form2fit: Learning shape priors for generalizable assembly from disassembly. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9404–9410. IEEE, 2020.
- [14] A. Ganapathi, P. Sundaresan, B. Thananjeyan, A. Balakrishna, D. Seita, J. Grannen, M. Hwang, R. Hoque, J. E. Gonzalez, N. Jamali, et al. Learning dense visual correspondences in simulation to smooth and fold real fabrics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11515–11522. IEEE, 2021.

- [15] M. Danielczuk, A. Kurenkov, A. Balakrishna, M. Matl, D. Wang, R. Martín-Martín, A. Garg, S. Savarese, and K. Goldberg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1614–1621. IEEE, 2019.
- [16] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.
- [17] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.
- [18] K. Burns, T. Yu, C. Finn, and K. Hausman. Robust manipulation with spatial features. In *CoRL 2022 Workshop on Pre-training Robot Learning*, 2022.
- [19] S. Belkhale, Y. Cui, and D. Sadigh. Data quality in imitation learning. *arXiv preprint arXiv:2306.02437*, 2023.
- [20] C. M. Barber, R. J. Shucksmith, B. MacDonald, and B. C. Wünsche. Sketch-based robot programming. In *2010 25th International Conference of Image and Vision Computing New Zealand*, pages 1–8. IEEE, 2010.
- [21] D. Porfirio, L. Stegner, M. Cakmak, A. Saupé, A. Albarghouthi, and B. Mutlu. Sketching robot programs on the fly. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI '23*, page 584–593, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399647. doi:10.1145/3568162.3576991. URL <https://doi.org/10.1145/3568162.3576991>.
- [22] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for Dynamics and Control Conference*, pages 893–905. PMLR, 2022.
- [23] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, et al. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches. *arXiv preprint arXiv:2311.01977*, 2023.
- [24] P. N. Chowdhury, A. K. Bhunia, A. Sain, S. Koley, T. Xiang, and Y.-Z. Song. What can human sketches do for object detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15083–15094, 2023.
- [25] A. K. Bhunia, S. Koley, A. Kumar, A. Sain, P. N. Chowdhury, T. Xiang, and Y.-Z. Song. Sketch2saliency: Learning to detect salient objects from human drawings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2733–2743, 2023.
- [26] A. K. Bhunia, V. R. Gajjala, S. Koley, R. Kundu, A. Sain, T. Xiang, and Y.-Z. Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2293–2302, 2022.
- [27] S. Qiu, S. Xie, L. Fan, T. Gao, J. Joo, S.-C. Zhu, and Y. Zhu. Emergent graphical conventions in a visual communication game. *Advances in Neural Information Processing Systems*, 35: 13119–13131, 2022.
- [28] Z. Lei, Y. Zhang, Y. Xiong, and S. Chen. Emergent communication in interactive sketch question answering. *arXiv preprint arXiv:2310.15597*, 2023.
- [29] P. N. Chowdhury, A. K. Bhunia, A. Sain, S. Koley, T. Xiang, and Y.-Z. Song. Scenetriology: On human scene-sketch and its complementarity with photo and text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10972–10983, 2023.

- [30] L. Zhang and M. Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.
- [31] S. Koley, A. K. Bhunia, A. Sain, P. N. Chowdhury, T. Xiang, and Y.-Z. Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6850–6861, 2023.
- [32] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.
- [33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [34] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermanno, D. Cohen-Or, A. Zamir, and A. Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.
- [35] Y. Vinker, Y. Alaluf, D. Cohen-Or, and A. Shamir. Clipascene: Scene sketching with different types and levels of abstraction. *arXiv preprint arXiv:2211.17256*, 2022.
- [36] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
- [37] M. Li, Z. Lin, R. Mech, E. Yumer, and D. Ramanan. Photo-sketching: Inferring contour drawings from images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1403–1412. IEEE, 2019.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [39] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [40] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba. Hindsight experience replay. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.
- [41] I. Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 1968.
- [42] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [43] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34: 12786–12797, 2021.
- [44] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.
- [45] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [46] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 1932.

A Additional Evaluations

In this section, we highlight the scale of our evaluations, additional findings from stress-testing RT-Sketch on sketches drawn by different individuals, and results from extending our policy to accommodate sketch+language conditioning.

A.1 Experiments At A Glance

Cumulatively, our results encompass the following: H1 experiments comprise 270 rollouts (6 skills x 15 trials x 3 methods), H2 comprises 40 rollouts (2 skills x 5 trials x 4 sketch types), H3 comprises 30 rollouts (15 trials x 2 methods), and H4 comprises 30 rollouts (15 trials x 2 methods). All rollouts are cumulatively evaluated across 62 labelers (split across H1-4).

A.2 Robustness to Input Sketches

To test whether RT-Sketch generalizes to sketches drawn by different individuals, we collect 30 *line sketches* (drawn via tracing) by 6 different annotators (whose sketches were never seen during training) on 5 trials of the *move near* scenario. We obtain the resulting rollouts produced by RT-Sketch with these sketches as input. Across ratings, RT-Sketch achieves high spatial alignment on sketches drawn by other annotators. Notably, the performance between sketches drawn by different annotators is similar, as well as the average across annotators compared to original policy performance on our original sketches (Fig. 4).

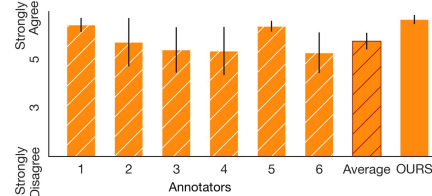


Figure 4: Sketches Drawn by Other Annotators

A.3 Multimodal Goal Specification: Sketches + Language

We train a sketch-and-language conditioned model by modifying the RT-1 architecture to use FiLM along with EfficientNet layers to tokenize both visual input and language, and concatenate them at the input. In H1 experiments (Fig. 3), we evaluate all policies on the *upright* skill, where the robot must place a can or bottle from a sideways orientation initially to an upright orientation at a desired location on the table. While RT-1 typically can reorient the can/bottle properly, it struggles to place the item in the intended location on the table, as reflected in this policy’s spatial imprecision in Table 1. Meanwhile, RT-Sketch struggles to reorient the can/bottle, since an imperfect sketch may fail to specify the exact desired orientation, but often places the can/bottle in the desired location. In Fig. 5, we see that while language alone (i.e. “place the can upright”) can be ambiguous in terms of spatial placement, and a sketch alone does not encourage reorientation, we empirically see that the joint policy is better able to address the limitations of either modality alone. A similar pattern emerges for *pick drawer* (Fig. 5).

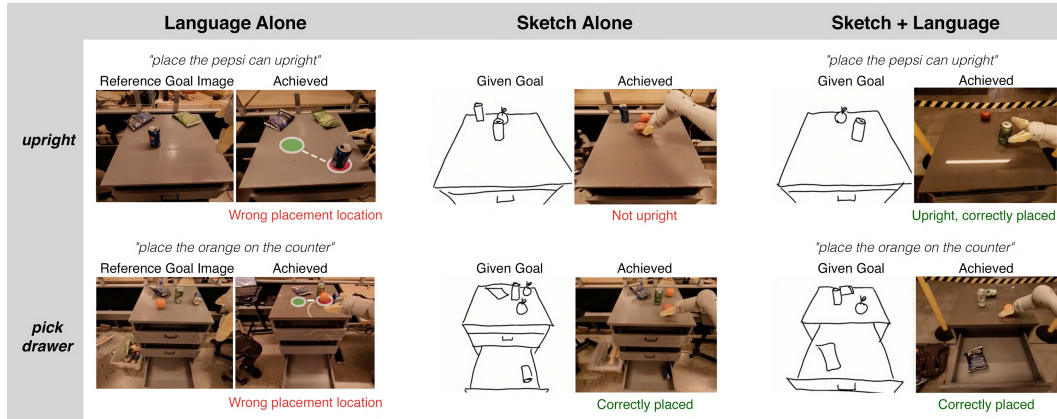


Figure 5: **Multimodal Goal Specification: Sketch+Language:** Empirically, we find that while a language-only policy can struggle with spatial precision, and a sketch-only policy can fail to interpret intended object orientations from a sketch alone, a multimodal policy is better able to address the limitations of both.

B Additional Results: Goal Alignment

In addition to the goal alignment results reported in Fig. 3 which are based on average Likert ratings, we additionally conduct a non-parametric Mann-Whitney U (MWU) test with $\alpha = 0.05$ for H1-4

to evaluate the differences in goal alignment ratings across methods. This kind of statistical test is suitable for ordinal data and does not make specific assumptions on the normality or variance of the data distributions.

B.1 H1 Findings

The H1 experiments aim to evaluate how RT-Sketch compares to RT-1 and RT-Goal-Image on the standard RT-1 tabletop manipulation benchmark [1]. We conduct a MWU test under the null hypothesis that there is no difference in the goal alignment ratings from labelers across the methods. In Appendix Table 3 and Appendix Table 4, we report the pairs of methods for which the ratings yield a p-value of < 0.05 , rejecting the null hypothesis, along with their U -statistic.

Table 3: **H1: RT-1 Benchmark - Semantic Alignment**

Skill	Method Pair	Stat.	p-value
Move Near			
Pick Drawer	(RT-1, RT-Goal Img)	5298.0	1.49×10^{-3}
Drawer Open	(RT-1, RT-Goal Img)	4797.0	1.22×10^{-3}
Drawer Close	(RT-1, RT-Goal Img)	4089.5	2.01×10^{-8}
Knock			
Upright	(RT-1, RT-Sketch)	16855.0	9.49×10^{-29}
	(RT-1, RT-Goal Img)	10052.0	2.80×10^{-18}
	(RT-Sketch, RT-Goal Img)	7210.5	5.62×10^{-7}

Table 4: **H1: RT-1 Benchmark - Spatial Alignment**

Skill	Method Pair	Stat.	p-value
Move Near			
Pick Drawer			
Drawer Open	(RT-1, RT-Goal Img)	4761.5	4.59×10^{-3}
Drawer Close	(RT-1, RT-Sketch)	7780.0	1.82×10^{-5}
	(RT-1, RT-Goal Img)	4869.0	3.62×10^{-10}
Knock			
Upright	(RT-1, RT-Sketch)	15085.0	1.55×10^{-14}
	(RT-1, RT-Goal Img)	10656.0	1.32×10^{-23}

We conclude that for 5 of 6 and 4 of 6 skills, the null hypothesis is confirmed for semantic and spatial alignment ratings, respectively, suggesting that there is no dropoff in performance with sketches compared to traditional modalities. We do observe that for the *upright* skill, the rating difference between RT-Sketch and RT-1 is significant, and RT-Sketch suffers a slight performance drop as re-orientation is particularly difficult to infer from a sketch alone. However, we have since addresses this challenge with a policy conditioned on both sketches and language, which performs reorientation better than sketches-alone and with more spatial precision than language-alone (Section 4.2).

The highlighted rows above indicate when the goal alignment ratings for RT-Sketch compared to either RT-1 or RT-Goal-Image were found to be statistically significant. Notably, there are very few such findings, in alignment with H1. This is in accordance with what we observe Fig. 3: nearly no noticeable difference in performance between methods for most of the skills, and the slightly better performance of RT-1 compared to RT-Sketch (and the slightly better performance of RT-Sketch compared to RT-Goal-Image) for the *upright* skill.

Table 5: **H2: Robustness to Sketch Specificity - Semantic Alignment**

Pair	Stat.	p-value
Free-Hand, Line Sketch	1059.0	9.58×10^{-12}
Free-Hand, Colored Sketch	960.0	2.54×10^{-10}
Free-Hand, Sobel Edges	1099.5	9.16×10^{-11}
Line Sketch, Colored Sketch	-	-
Line Sketch, Sobel Edges	-	-
Colored Sketch, Sobel Edges	-	-

Table 6: **H2: Robustness to Sketch Specificity - Spatial Alignment**

Pair	Stat.	p-value
Free-Hand, Line Sketch	478.0	5.18×10^{-17}
Free-Hand, Colored Sketch	567.5	3.49×10^{-13}
Free-Hand, Sobel Edges	629.0	3.09×10^{-14}
Line Sketch, Colored Sketch	-	-
Line Sketch, Sobel Edges	-	-
Colored Sketch, Sobel Edges	-	-

B.2 H2 Findings

For H2 experiments, we evaluate RT-Sketch’s robustness to the input specificity of the sketch. We find that across the 4 sketch types, the only pairings which garner statistically significant differences in ratings are free-hand sketches as compared to other types (Appendix Table 5 and Appendix Table 6). This is natural given the drastic perspective and geometric differences of free-hand sketches compared to those which are *traced* or derived from a transform of the goal image itself (edge detection).

However, there are notably no statistically significant pairings between line-sketches and even the most detailed type of input representation we evaluate (Sobel Edges). This suggests that RT-Sketch is indeed able to handle a range of input specificity levels, and more importantly that RT-Sketch can deal with representations that are minimal and imperfect.

Table 7: **H3: Visual Distractors**

Alignment	Method Pair	Stat.	p-value
Semantic	RT-Sketch, RT-Goal Img.	20622.5	4.62×10^{-8}
Spatial	RT-Sketch, RT-Goal Img.	22233.0	3.07×10^{-12}

Table 8: **H4: Language Ambiguity**

Alignment	Method Pair	Stat.	p-value
Semantic	RT-Sketch, RT-1	4756.0	1.34×10^{-24}
Spatial	RT-Sketch, RT-1	3680.5	3.53×10^{-30}

B.3 H3 and H4 Findings

Finally, we conduct a MWU test over the semantic/spatial goal alignment ratings between RT-Sketch and RT-Goal-Image in the setting of visual distractors (H3, Appendix Table 7) as well as RT-Sketch and RT-1 in the setting of language ambiguity (H4, Appendix Table 8). We hypothesize that RT-Sketch does indeed achieve *higher* ratings than baselines in these settings, as sketches are by nature 1) minimal, which may enable emergent robustness to distractors, and 2) agnostic to language.

We do find a statistically significant difference across semantic and spatial ratings (highlighted in orange), concluding that RT-Sketch is favorable to traditional modalities in these particular settings.

B.4 Summary of Mann-Whitney U Findings

In short, the additional findings from conducting more thorough MWU testing over H1-4 align very closely with what we observe and report in Fig. 3 and suggest the merits of sketches across a range of scenarios.

C Future Directions

Learning a policy conditioned on view-invariant sketches can be an initial step before moving to even more abstract representations like schematics or diagrams for assembly tasks. Additionally, alternative ways to condition on sketches is a powerful avenue for future work. RT-Sketch currently only considers goal observations in sketch space, but projecting all observations to a sketch-based

or latent space is another underexplored but promising direction. Sketches are not without their own limitations, however, as ambiguity due to omitted details or poor quality sketches are persistent challenges. In the future, we are excited to continue exploring multimodal goal specification which can leverage the benefits of language, sketches, and other modalities to jointly resolve ambiguity from any single modality alone. This may include both end-to-end approaches that can jointly condition on multiple modalities, or hierarchical strategies that can leverage the spatial awareness of sketches and the summarization capabilities of VLMs to supplement ambiguous language with more informed descriptions derived from visual observations of a sketch. Lastly, exploring what combination of modalities humans prefer to use when providing goals, and how best they capture intent, is an important future direction not addressed in this work.

D Sketch Goal Representations

Since the main bottleneck to training a sketch-to-action policy like RT-Sketch is collecting a dataset of paired trajectories and goal sketches, we first train an image-to-sketch translation network \mathcal{T} mapping image observations o_i to sketch representations g_i , discussed in Section 3. To train \mathcal{T} , we first take a pre-trained network for sketch-to-image translation [37] trained on the ContourDrawing dataset of paired images and edge-aligned sketches (Fig. 6). This dataset contains $L^{(i)} = 5$ crowd-sourced sketches per image for 1000 images. By pre-training on this dataset, we hope to embed a strong prior in \mathcal{T} and accelerate learning on our much smaller dataset. Next, we finetune \mathcal{T} on a dataset of 500 manually drawn line sketches for RT-1 robot images. We visualize a few examples of our manually sketched goals in Fig. 7 under ‘Line Drawings’.

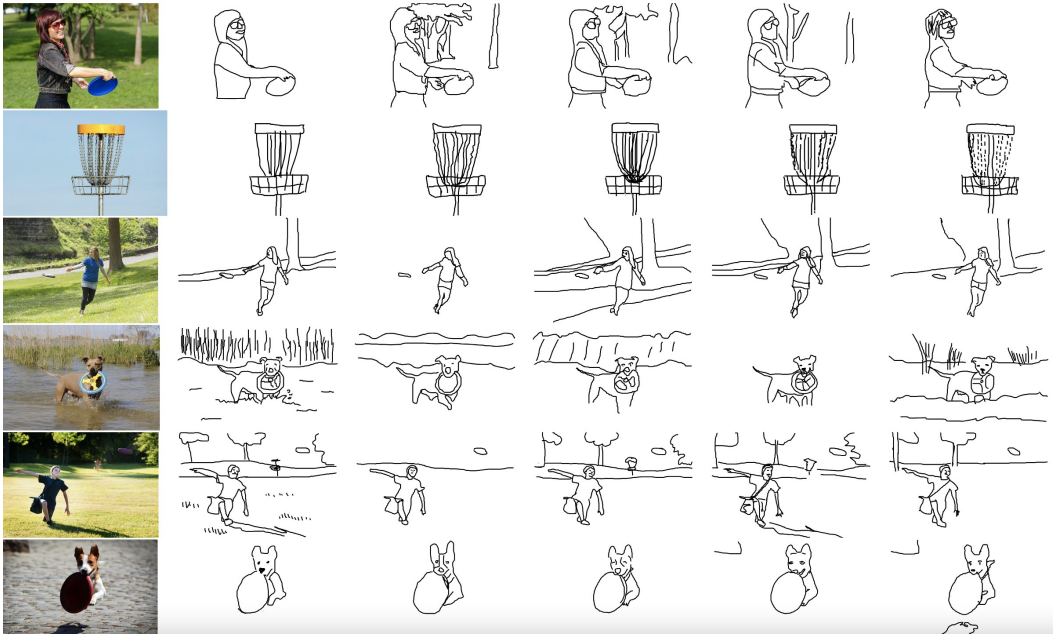


Figure 6: **ContourDrawing Dataset:** We visualize 6 samples from the ContourDrawing Dataset from [37]. For each image, 5 separate annotators provide an edge-aligned sketch of the scene by outlining on top of the original image. As depicted, annotators are encouraged to preserve main contours of the scene, but background details or fine-grained geometric details are often omitted. Li et al. [37] then train an image-to-sketch translation network \mathcal{T} with a loss that encourages aligning with at least one of the given reference sketches.

Notably, while we only train \mathcal{T} to map an image to a black-and-white line sketch \hat{g}_i , we consider various augmentations \mathcal{A} on top of generated goals to simulate sketches with varied colors, affine and perspective distortions, and levels of detail. Fig. 7 visualizes a few of these augmentations, such as automatically colorizing black-and-white sketches by superimposing a blurred version of the original RGB image, and treating an edge-detected version of the original image as a generated sketch to simulate sketches with a lot of details. We generate a dataset for training RT-Sketch by ‘sketchifying’ hind-sight relabeled goal images via \mathcal{T} and \mathcal{A} .

Although RT-Sketch is only trained on generated line sketches, colorized line sketches, edge-detected images, and goal images, we find that it is able to handle sketches of even greater diversity.

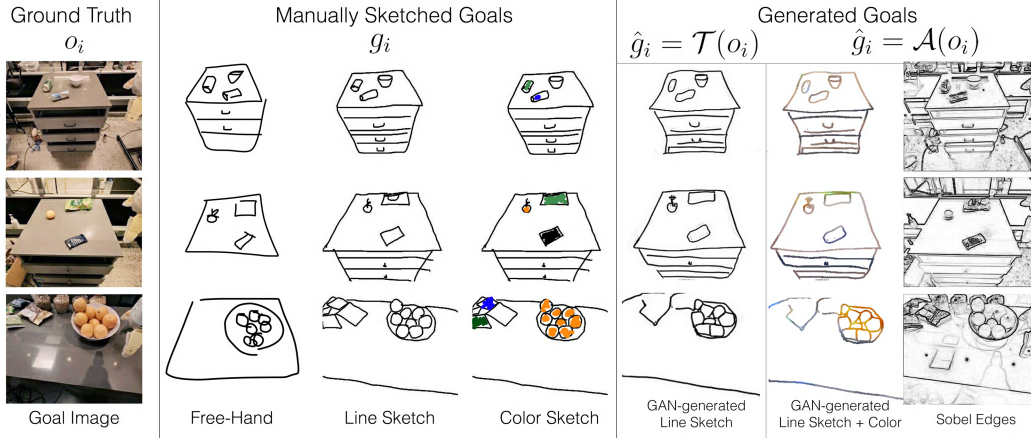


Figure 7: **Visual Goal Diversity**: RT-Sketch is capable of handling a variety of visual goals at both train and test time. RT-Sketch is trained on generated and augmented images like those shown on the right below ‘Generated Goals’. But it can also interpret free-hand, line sketches, and colored sketches at test time such as those on the left below ‘Manually Sketched Goals’.

This includes non-edge aligned free-hand sketches and sketches with color infills, like those shown in Fig. 7.

D.1 Alternate Image-to-Sketch Techniques

The choice of image-to-sketch technique we use is critical to the overall success of the RT-Sketch pipeline. We experiment with various other techniques before converging on the above approach.

Recently, two recent works, CLIPAsso [34] and CLIPAScene [35] explore methods for automatically generating a sketch from an image. These works pose sketch generation as inferring the parameters of Bezier curves representing “strokes” in order to produce a generated sketch with maximal CLIP-similarity to a given input image. These methods perform a per-image optimization to generate a plausible sketch, rather than a global batched operation across many images, limiting their scalability. Additionally, they are fundamentally more concerned with producing high-quality, aesthetically pleasing sketches which capture a lot of extraneous details.

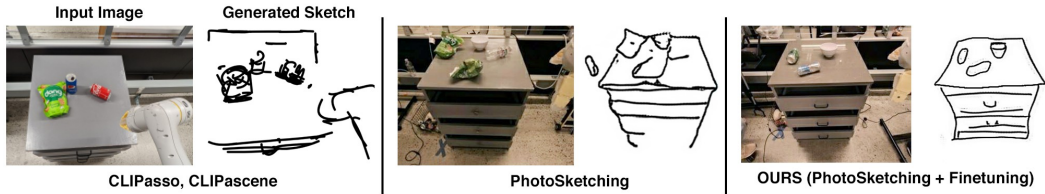


Figure 8: **Alternate Image-to-Sketch Techniques**

We, on the other hand, care about producing a minimal but reasonable-quality sketch. The second technique we explore is trying the pre-trained Photosketching GAN [37] on internet data of paired images and sketches. However, this model output does not capture object details well, likely due to not having been trained on robot observations, and contains irrelevant sketch details. Finally, by finetuning this PhotoSketching GAN on our own data, the outputs are much closer to real, hand-drawn human sketches that capture salient object details as minimally as possible. We visualize these differences in Fig. 8.

E Evaluation Visualizations

To further interpret RT-Sketch’s performance, we provide visualizations of the precision metrics and experimental rollouts. In Fig. 9, we visualize the degree of alignment RT-Sketch achieves, as quantified by the pixelwise distance of object centroids in achieved vs. given goal images. In Fig. 10, Fig. 11, Fig. 12, and Fig. 14, we visualize each policy’s behavior for H1, H2, H3 and H4, respectively. Fig. 13 visualizes the four tiers of difficulty in language ambiguity that we analyze for H4.

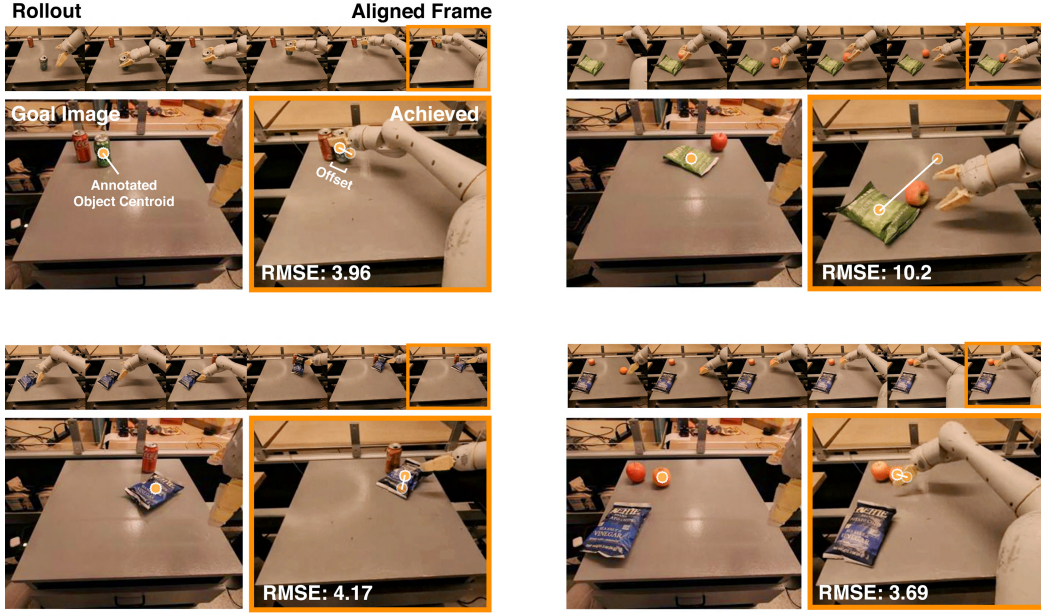


Figure 9: **Spatial Precision Visualization:** We visualize four trials of RT-Sketch on the Move Near skill, along with the measured spatial precision in terms of RMSE. To evaluate spatial precision, we have a human annotator annotate the frame that is visually most aligned, and then keypoints for the object that was moved in this frame and in the provided reference goal image. For each of the four trials, we visualize the rollout frames until alignment is achieved, along with the labeled object centroids and the offset in achieved vs. desired positions. The upper right example shows a failure of RT-Sketch in which the apple is moved instead of the chip bag, incurring a high RMSE. These visualizations are intended to better contextualize the numbers from Table 1.

F RT-Sketch Failure Modes and Limitations

While RT-Sketch is performant at several manipulation benchmark skills, capable of handling different levels of sketch detail, robust to visual distractors, and unaffected by ambiguous language, it is not without failures and limitations.

In Fig. 16, we visualize the failure modes of RT-Sketch. One failure mode we see with RT-Sketch is occasionally re-trying excessively, as a result of trying to align the scene as closely as possible. For instance, in the top row, Rollout Image 3, the scene is already well-aligned, but RT-Sketch keeps shifting the chip bag which causes some misalignment in terms of the chip bag orientation. Still, this kind of failure is most common with RT-Goal-Image (Table 1), and is not nearly as frequent for RT-Sketch. We posit that this could be due to the fact that sketches enable high-level spatial reasoning without over-attending to pixel-level details.

One consequence of spatial reasoning at such a high level, though, is an occasional lack of precision. This is noticeable when RT-Sketch orients items incorrectly (second row) or positions them slightly off, possibly disturbing other items in the scene (third row). This may be due to the fact that sketches are inherently imperfect, which makes it difficult to reason with such high precision.

Finally, we see that RT-Sketch occasionally manipulates the wrong object (rows 4 and 5). Interestingly, we see that a fairly frequent pattern of behavior is to manipulate the wrong object (orange in row 4) to the right target location (near green can in row 4). This may be due to the fact that the sketch-generating GAN has occasionally hallucinated artifacts or geometric details missing from the actual objects. Having been trained on some examples like these, RT-Sketch can mistakenly perceive the wrong object to be aligned with an object drawn in the sketch. However, the sketch still indicates the relative desired spatial positioning of objects in the scene, so in this case RT-Sketch still attempts to align the incorrect object with the proper place.

Finally, the least frequent failure mode is manipulating the wrong object to the wrong target location (i.e. opening the wrong drawer handle). This is most frequent when the input is a free-hand sketch, and could be mitigated by increasing sketch detail (Table 2).

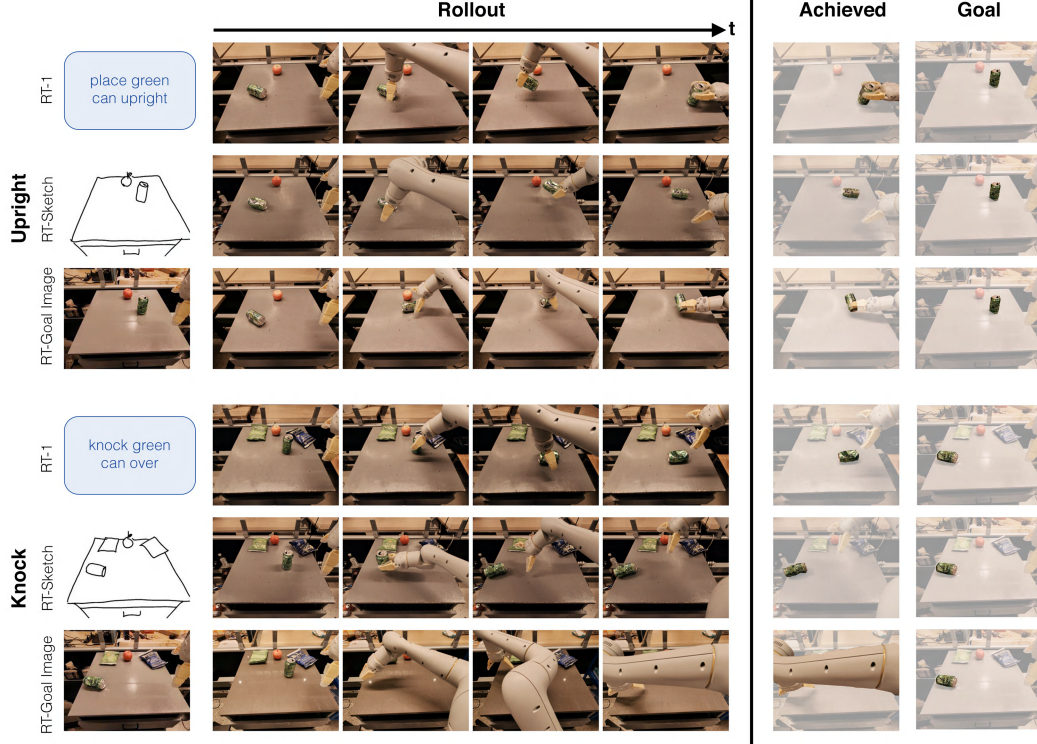


Figure 10: **H1 Rollout Visualization:** We visualize the performance of RT-1, RT-Sketch, and RT-Goal-Image on two skills from the RT-1 benchmark (*upright* and *knock*). For each skill, we visualize the goal provided as input to each policy, along with the policy rollout. We see that for both skills, RT-1 obeys the semantic task at hand by successfully placing the can upright or sideways, as intended. Meanwhile, RT-Sketch and RT-Goal-Image struggle with orienting the can upright, but successfully knock it sideways. Interestingly, both RT-Sketch and RT-Goal-Image are able to place the can in the desired location (disregarding can orientation) whereas RT-1 does not pay attention to where in the scene the can should be placed. This is indicated by the discrepancy in position of the can in the achieved versus goal images on the right. This trend best explains the anomalous performance of RT-Sketch and RT-Goal-Image in perceived Likert ratings for the upright task (Fig. 3), but validates their comparably higher spatial precision compared to RT-1 across all benchmark skills (Table 1).

660 G Evaluation and Assessment Interfaces

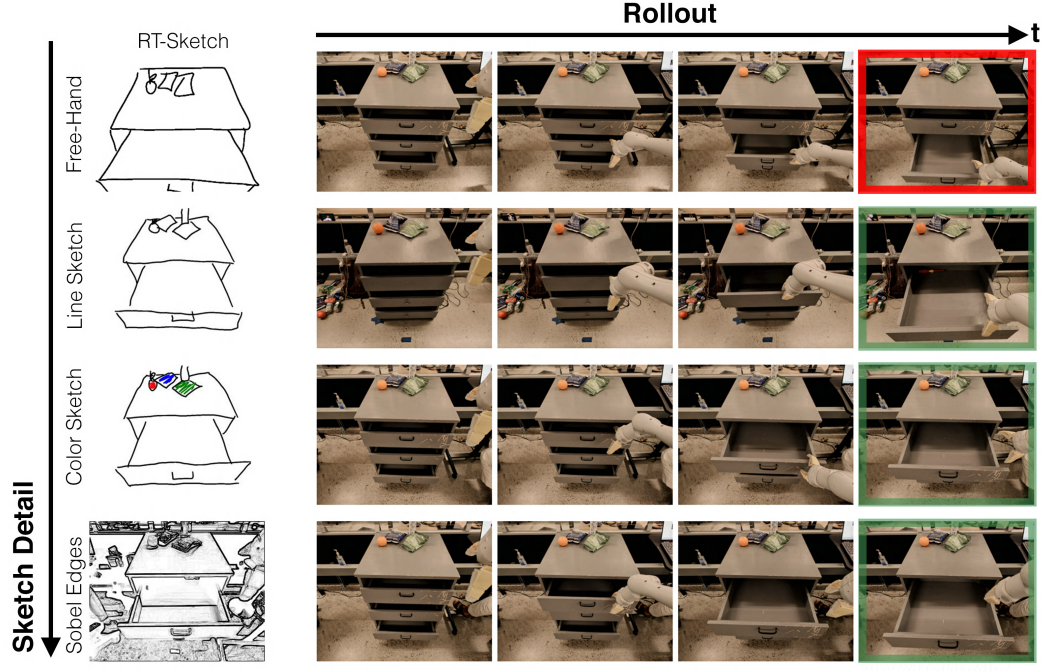


Figure 11: **H2 Rollout Visualization:** For the *open drawer* skill, we visualize four separate rollouts of RT-Sketch operating from different input types. Free-hand sketches are drawn without outlining over the original image, such that they can contain marked perspective differences, partially obscured objects (drawer handle), and roughly drawn object outlines. Line sketches are drawn on top of the original image using the sketching interface we present in Appendix Fig. 17. Color sketches merely add color infills to the previous modality, and Sobel Edges represent an upper bound in terms of unrealistic sketch detail. We see that RT-Sketch is able to successfully open the correct drawer for any sketch input except the free-hand sketch, without a noticeable performance gain or drop. For the free-hand sketch, RT-Sketch still recognizes the need for opening a drawer, but the differences in sketch perspective and scale can occasionally cause the policy to attend to the wrong drawer, as depicted.

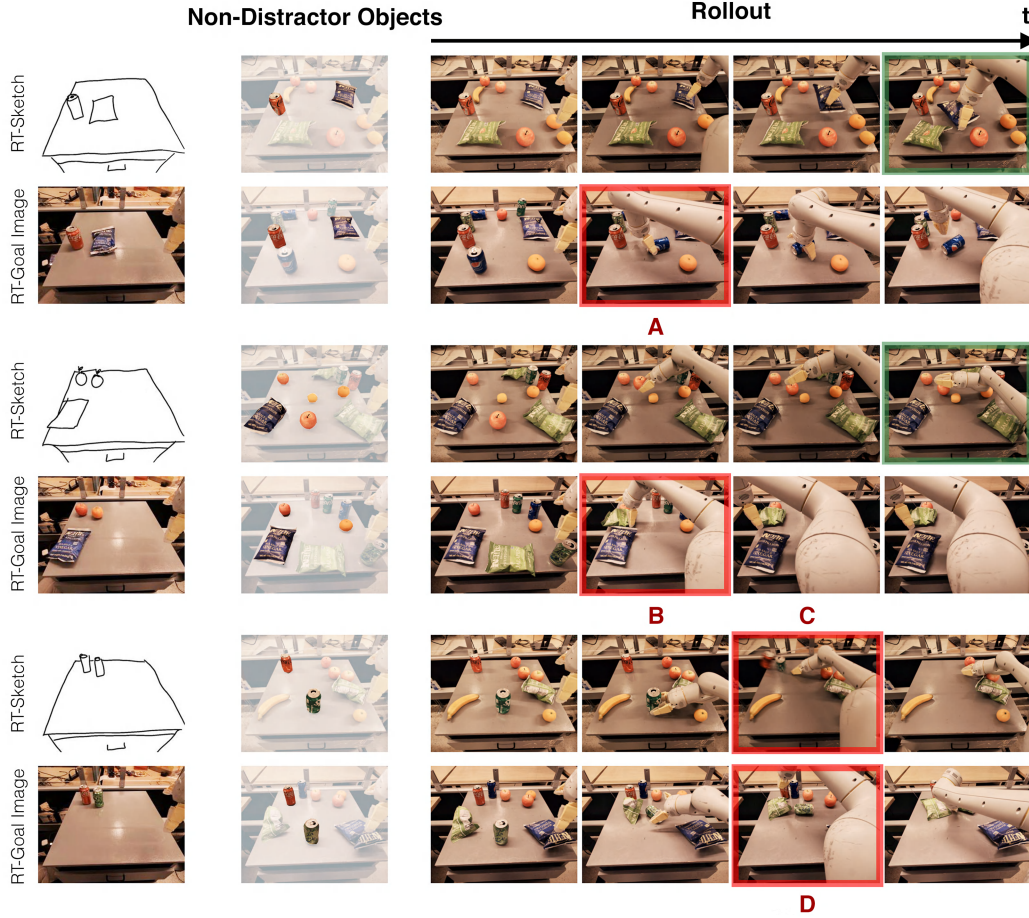


Figure 12: **H3 Rollout Visualization:** We visualize qualitative rollouts for RT-Sketch and RT-Goal-Image for 3 separate trials of the *move near* skill subject to distractor objects. In Column 2, we highlight the relevant non-distractor objects that the policy must manipulate in order to achieve the given goal. In Trial 1, we see that RT-Sketch successfully attends to the relevant objects and moves the blue chip bag near the coke can. Meanwhile, RT-Goal-Image is confused about which blue object to manipulate, and picks up the blue pepsi can instead of the blue chip bag (A). In Trial 2, RT-Sketch successfully moves an apple near the fruit on the left. A benefit of sketches is their ability to capture instance multimodality, as any of the fruits highlighted in Column 2 are valid options to move, whereas this does not hold for an overspecified goal image. RT-Goal-Image erroneously picks up the green chip bag (B) instead of a fruit. Finally, Trial 3 shows a failure for both policies. While RT-Sketch successfully infers that the green can must be moved near the red one, it accidentally knocks over the red can (C) in the process. Meanwhile, RT-Goal-Image prematurely drops the green can and instead tries to pick the green chip bag (D).

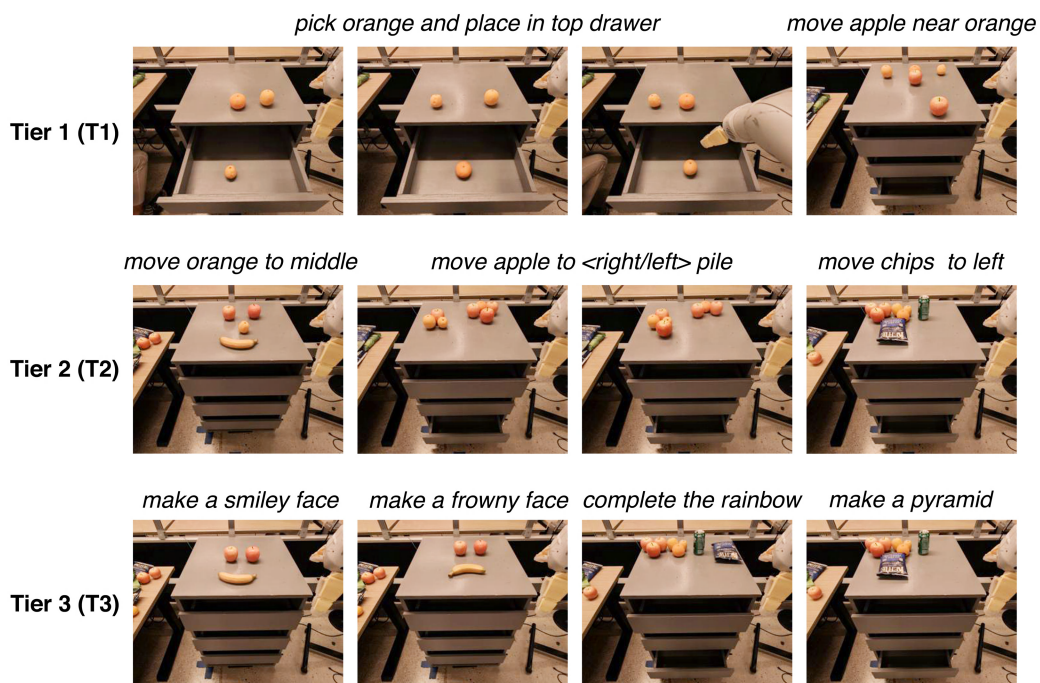


Figure 13: **H4 Tiers of Difficulty**: To test **H4**, we consider language instructions that are either ambiguous due the presence of multiple similar object instances (**T1**), are somewhat out-of-distribution for RT-1 (**T2**), or are far out-of-distribution and difficult to specify concretely without lengthier descriptions (**T3**). Each image represents the ground truth goal image paired with the task description.

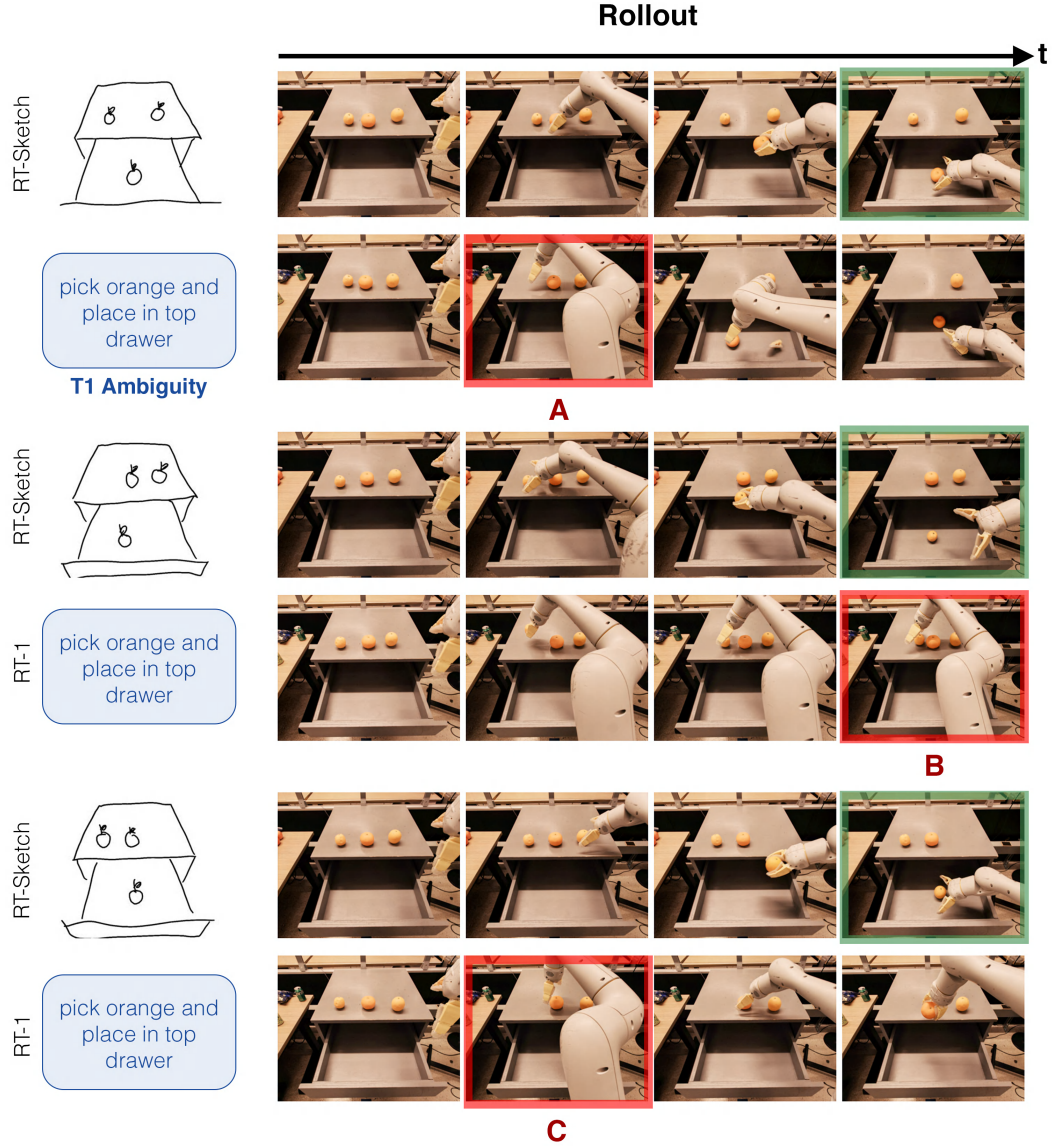


Figure 14: **H4 Rollout Visualization (T1 as visualized in Fig. 13)**: One source of ambiguity in language descriptions is mentioning an object for which there are multiple instances present. For example, we can easily illustrate three different desired placements of an orange in the drawer via a sketch, but an ambiguous instruction cannot easily specify which orange is relevant to pick and place. In all rollouts, RT-Sketch successfully places the correct orange in the drawer, while RT-1 either picks up the wrong object (A), fails to move to the place location (B), or knocks off one of the oranges (C). Although in this case, the correct orange to manipulate could easily be specified with a spatial relation like *pick up the $\langle \text{left/middle/right} \rangle$ orange*, we show below in Appendix Fig. 15 that this type of language is still out of the realm of RT-1’s semantic familiarity.

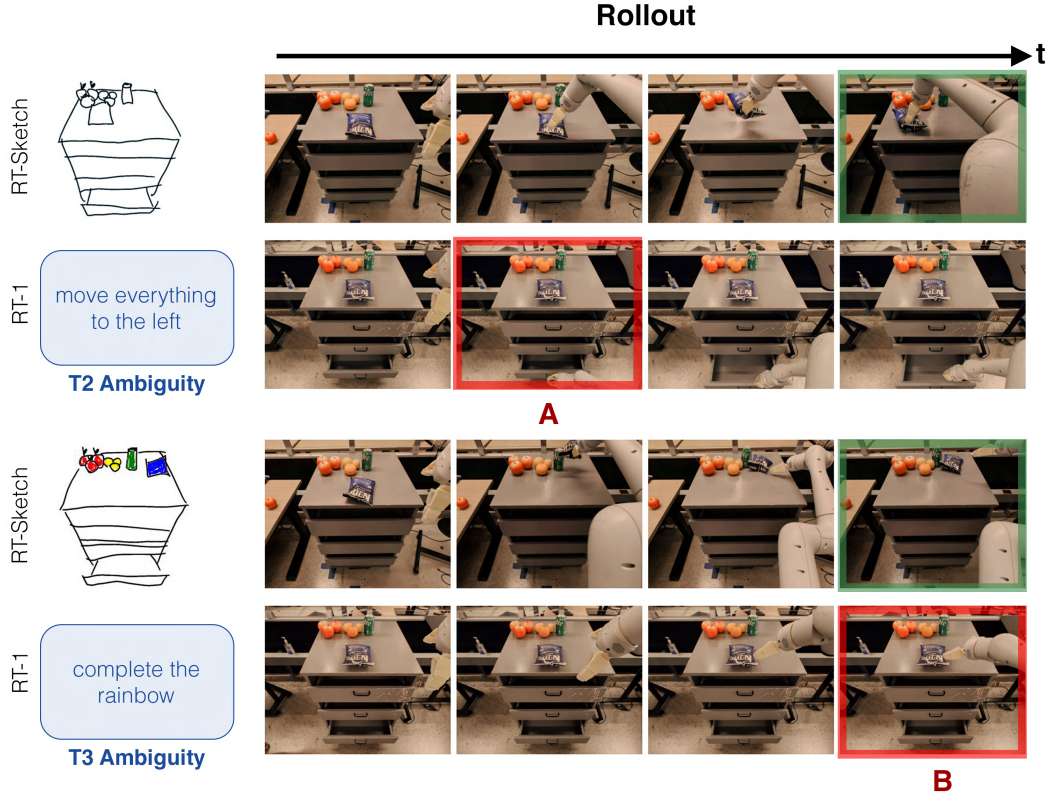


Figure 15: **H4 Rollout Visualization (T2-3 as visualized in Fig. 13)**: For **T2**, we consider language with spatial cues that intuitively should help the policy disambiguate in scenarios like the oranges in Fig. 14. However, we find that RT-1 is not trained to handle such spatial references, and this kind of language causes a large distribution shift leading to unwanted behavior. Thus, for the top rollout of trying to move the chip bag to the left where there is an existing pile, RT-Sketch completes the skill without issues, but RT-1 attempts to open the drawer instead of even attempting to rearrange anything on the countertop (A). For **T3**, we consider language goals that are even more abstract in interpretation, without explicit objects mentioned or spatial cues. Here, sketches are advantageous in their ability to succinctly communicate goals (i.e. visual representation of a rainbow), whereas the corresponding language task string is far too underspecified and OOD for the policy to handle (B).

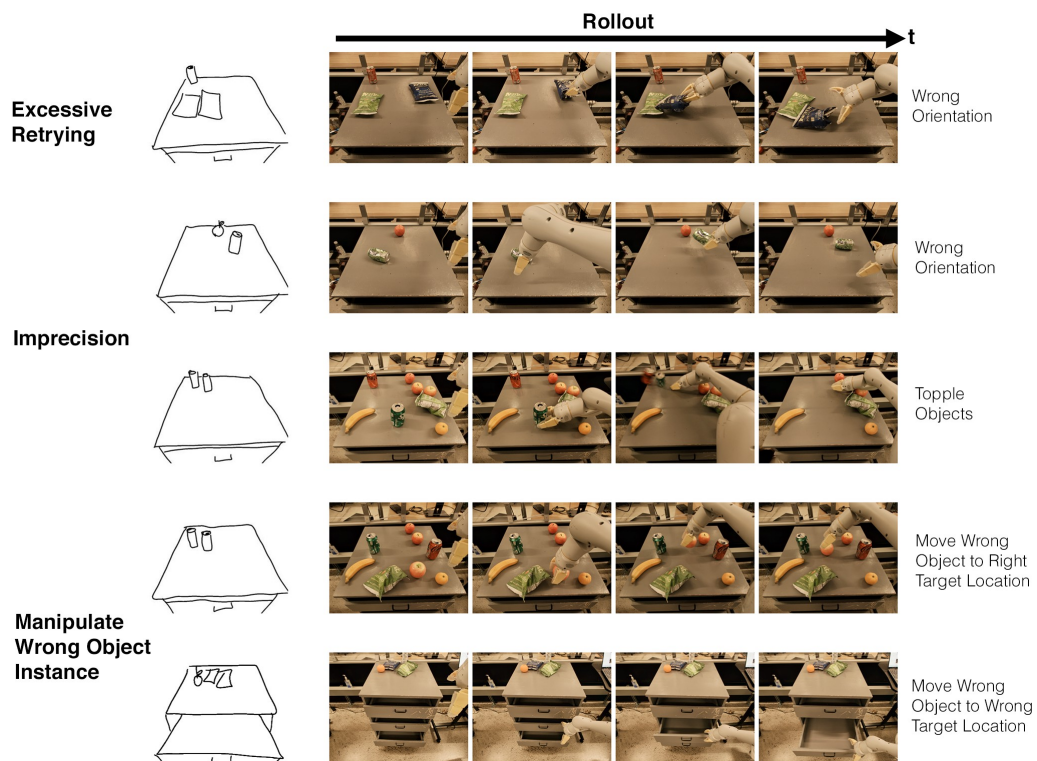


Figure 16: RT-Sketch Failure Modes

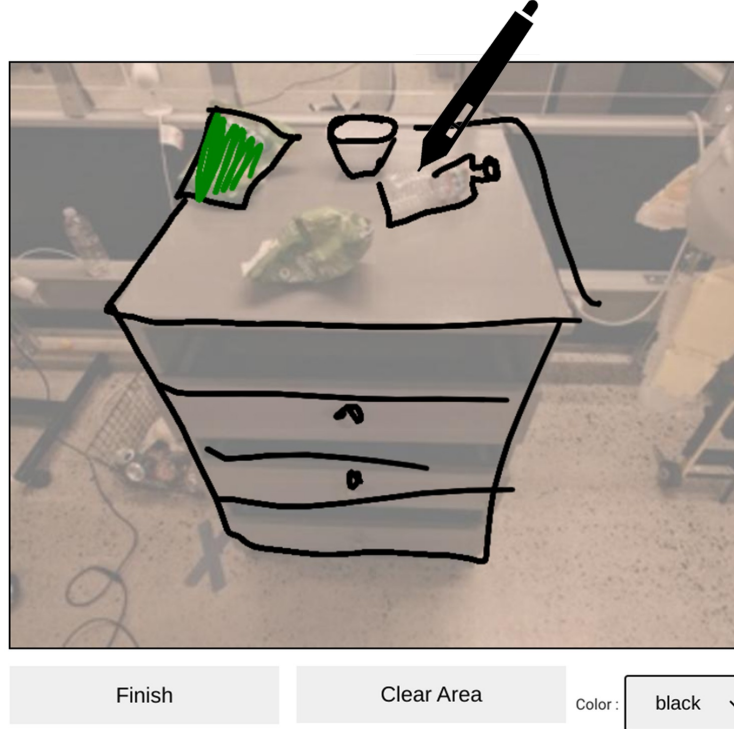


Figure 17: **Sketching UI:** We design a custom sketching interface for manually collecting paired robot images and sketches with which to train \mathcal{T} , and for sketching goals for evaluation. The interface visualizes the current robot observation, and provides the ability to draw on a digital screen with a stylus. The above visualization shows the color-sketching modality, which is a traced representation with color shading. The interface supports different colors and erasure, along with either *tracing* over the image (line-sketching) or drawing free-form over a blank canvas (free-hand sketches). We note that intuitively, drawing on top of the image is not an unreasonable assumption to make, since current agent observations are typically readily available compared to a goal image, for instance. Additionally, the overlay is intended to make the sketching interface easy for the user to provide, without having to eyeball edges for the drawers or handles blindly. This provides helpful guides for sketching and is an easy way to obtain sketches that more closely align with current observations for free.

Q1

Reference Instruction + Actual Rollout



The robot achieves **semantic alignment** with the goal during the rollout. *

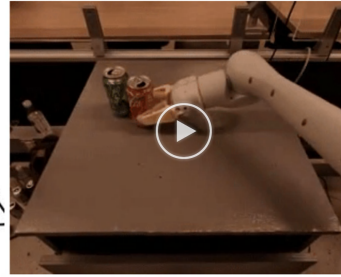
	1	2	3	4	5	6	7	
Strongly Disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly Agree

Q2

Reference Goal



Actual Rollout



The robot achieves **spatial alignment** with the goal during the rollout. *

	1	2	3	4	5	6	7	
Strongly Disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly Agree

Figure 18: **Assessment UI**: For all skills and methods, we ask labelers to assess semantic and spatial alignment of the recorded rollout relative to the ground truth semantic instruction and visual goal. We show the interface above, where labelers are randomly assigned to skills and methods (anonymized). The results of these surveys are reported in [Fig. 3](#).