マルチモーダルLLM及び視覚言語基盤モデルに基づく多階層アラインメント表現による物体操作タスク成功判定

○後神美結，神原元就，小槻誠太郎，杉浦孔明（慶應義塾大学）

本研究では，open-vocabulary 物体操作タスクの成功を，物体操作前後の一人称視点画像と指示文に基づいて予測するタスクを扱う．従来のマルチモーダル大規模言語モデル（MLLM）を含む手法は，物体の詳細な特徴や位置の微細な変化を十分に理解できないことがある．そこで我々は，2 つの画像間の視覚的表現の違いに注目し，重要な変化に着目する Contrastive λ-Repformer を提案する．大規模標準データセットである RT-1 データセットおよび実世界のロボットプラットフォームを用いた実験の結果，提案手法が代表的な MLLM ベースのモデルの精度を 13.52 ポイント上回った．

1. はじめに

物体操作におけるタスク成功予測は，正確で効率的な操作を可能にし，医療，産業，農業，環境管理，物流などのロボット応用において，信頼性と一貫性を向上させ得る．例えば，産業における部品の組み立てや，農業における作物の収穫等の物体操作タスクにおいて，マニピュレータがタスク成功を予測できると，品質や生産性を向上させることができる．サブタスクの失敗が後続のタスクに影響を及ぼす可能性があるため，サブタスクの成否を正確に予測することは，長期的なタスクにとって，特に重要である．

タスクの代表例を図 1 に示す．図 1 では，指示文 “pick a red can in the front right” に対応する物体操作の実行前後に撮影された 2 枚の一人称視点画像が示されている．このケースでは，マニピュレータは指示文通りに物体操作を実行したため，モデルは成功と予測することが期待される．

タスクは以下の 2 点により困難である．まず，物体操作前後に撮影された画像の変化，画像内の物体に関する情報，open-vocabulary 物体操作の指示文に関する十分な理解が必要な点である．また，このタスクでは，上記の要素が互いにアラインしているかの判断が求められる．6 節で示すように，マルチモーダル大規模言語モデル（MLLM [1–3]）をこのタスクに用いた場合，性能が限定的である．これは，MLLM が物体の詳細な特徴や位置の微細な変化を十分に理解できないことが多いためである．

本研究では，画像と指示文の間のアラインメントを行うことで，open-vocabulary 物体操作タスクの成否を予測する Contrastive λ-Repformer を提案する．本手法は，λ-Representation [4] を拡張した視覚表現に基づいて，タスク成功判定を行う．さらに，提案手法は画像間の差異の表現を用いることで，指示文とこの差異を効果的にアラインすることができる．これにより，モデルは物体の詳細な特徴や，それらの空間的関係を考慮した指示文の理解が可能となる．

本研究の新規性は以下の通りである．

- 2 つの画像の λ-Representation の違いを抽出する Contrastive λ-Representation Decoder を提案する．このモジュールを用いることでタスク成功予測を行う際に，画像間の差異と指示文とのアラインメントを考慮することができる．
- λ-Representation を拡張した視覚表現に基づく視覚特徴を導入する．

図 1 対象タスクに含まれるサンプルの例．本タスクは，物体操作前後に撮影された一人称視点画像と指示文に基づいて物体操作の成否を予測することである．

2. 関連研究

ロボティクス分野において，基盤モデルの応用に関する研究は増加している [5–7]．複数のサーベイ [8, 9] で本分野での MLLM ベースモデルが包括的に要約されている．本分野のマルチモーダル言語理解タスクにおいて，様々なデータセットが実世界環境 [10–12] およびシミュレーション環境 [12–14] で利用されている．

物体操作タスクにおいて，大規模言語モデル（LLM）はタスクプランニングに頻繁に使用される [12, 15, 16]．例として，高次の指示文からサブゴールを生成するために LLM を利用する研究が挙げられる [15, 16]．我々の手法は MLLM ベースのタスクプランニングモデル [16, 17] とも密接に関連している．従来手法とは異なり，提案手法は自然言語を考慮して画像を構造化する目的で MLLM を用いる．さらに，本手法では，2 種類のモジュールによって視覚表現を抽出し，それらを MLLM とともに統合するメカニズムを導入する．これにより，本手法は単純な MLLM ベースのアプローチでは十分に捉えることができない，多階層アラインメントを行った視覚表現を考慮することが可能になる．

物体操作中の衝突予測タスクも我々のタスクに関連がある．このタスクに対する手法として，事後に衝突判定を行う手法 [18] や，画像と配置ポリシーから衝突を予測する手法がある [19, 20]．我々の手法は，衝突以外のタスクの失敗に寄与する要因も考慮できる点において，これらの手法と異なる．

3. 問題設定

本研究では，Success Prediction for Open-vocabulary Manipulation（SPOM）タスクを扱う．SPOM とは，物体操作前後に撮影された一人称視

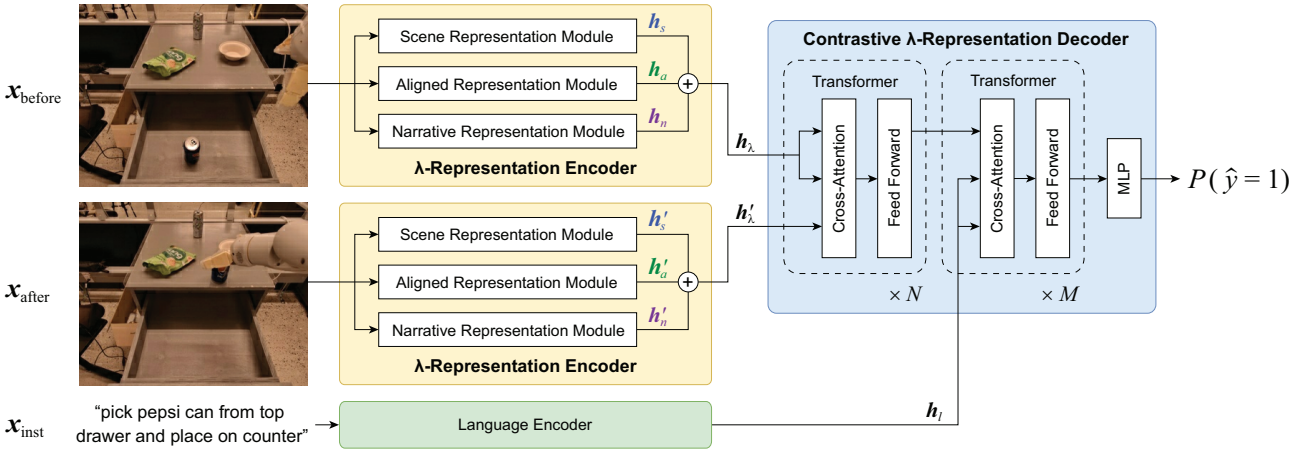


図 2 Contrastive λ-Repformer の概要．指示文と操作前後の画像が与えられたとき，本モデルはマニピュレータが操作を成功させたと予測される確率 $P(\hat{y}=1)$ を出力する．

点画像と指示文が与えられたときに open-vocabulary 物体操作タスクが成功したか否かを予測するタスクである．本タスクにおいてモデルは物体操作の成功または失敗を適切に予測することが期待される．

入力は，物体操作前と物体操作後に撮影された一人称視点画像および指示文から構成される．出力は予測確率 $P(\hat{y}=1)$ であり，これは指示文に基づく物体操作が適切に実行された確率の予測値を示す．ただし，$\hat{y}$ は物体操作の成否を示す予測ラベルであり，成功を '1' とする．本研究では，入力画像として一人称視点画像のみを使用する．また，一部の画像では，シーンがマニピュレータによって部分的に遮られている．タスクは実行可能であるものの，対象物やエリアが部分的に遮られることでタスク実行が難しいサンプルも存在する．

4. 提案手法

図 2 に提案手法である Contrastive λ-Repformer の構造を示す．入力は $\boldsymbol{x} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{before}}, \boldsymbol{x}_{\text{after}}\}$ である．ここで，$\boldsymbol{x}_{\text{inst}}$ はトークン化された指示文，$\boldsymbol{x}_{\text{before}}$ および $\boldsymbol{x}_{\text{after}}$ はそれぞれ物体操作の前後に撮影された RGB 画像を示している．提案手法の主なモジュールは λ-Representation Encoder および Contrastive λ-Representation Decoder である．

4.1 λ-Representation Encoder

λ-Representation [4] を拡張した視覚表現を抽出するための λ-Representation Encoder を導入する．このモジュールでは，3 種類の視覚表現を得て，それらを λ-Representation として統合する．図 2 に示されているように，このモジュールは Scene Representation Module，Aligned Representation Module，および Narrative Representation Module の 3 つのサブモジュールから構成される．λ-Representation Encoder の入力は $\boldsymbol{x}_{\text{before}}$ または $\boldsymbol{x}_{\text{after}}$ である．最初に，Language Encoder を用いて $\boldsymbol{x}_{\text{inst}}$ から言語特徴量 $\boldsymbol{h}_l$ を抽出する．

次に，Scene Representation $\boldsymbol{h}_s = f_{\text{srm}}(\boldsymbol{x}_{\text{before}})$ を取得する．ただし，$f_{\text{srm}}(\cdot)$ は Scene Representation Module を示している．Scene Representation Module は複数のバックボーンネットワークで構成される．つづいて，Aligned Representation $\boldsymbol{h}_a$ はマルチモーダル基盤モデルからなる Aligned Representation Module を用いて取得する．この特徴量は自然言語と適切にアラインメントがとれているため，Aligned Representation として扱うことができる．Narrative Representation $\boldsymbol{h}_n$ は MLLM と複数のテキスト埋め込みモデルによって構成される Narrative Representation Module を使って取得する．テキストプロンプトは物体の色，大きさ，形状，配置方法，画像内での位置，および他の物体との相対位置に着目するように設計した．これらの特徴量を結合し，$\boldsymbol{h}_n$ とする．最後に，$\boldsymbol{x}_{\text{before}}$ の λ-Representation $\boldsymbol{h}_\lambda = \left[\boldsymbol{h}_s^\mathsf{T}, \boldsymbol{h}_a^\mathsf{T}, \boldsymbol{h}_n^\mathsf{T}\right]^\mathsf{T}$ を取得する．同様にして，$\boldsymbol{x}_{\text{after}}$ の λ-Representation $\boldsymbol{h}'_\lambda$ も得る．

4.2 Contrastive λ-Representation Decoder

$\boldsymbol{h}_\lambda$ と $\boldsymbol{h}'_\lambda$ 間の表現の差異を獲得するために Contrastive λ-Representation Decoder を提案する．物体操作による物体の変化は画像間の差分に含まれる．従って，この表現を用いることでマニピュレータの物体操作に起因する可能性がある差異に焦点を当てることができる．一方で，画像間の差分は，必ずしも指示文によって指定されたタスクの成功を表すわけではない．例えば，Fig. 2 の例において，ペプシの缶が倒れてしまっていた場合，2 枚の画像の間には差分が存在するものの，タスクは失敗したと言える．従って，画像間の差分のみでタスクの成否を判定することは困難である．ゆえに，タスクの成否を判定する際は，表現の差分および指示文の間のアラインメントを考慮することが重要である．

このモジュールの入力は $\boldsymbol{h}_\lambda$，$\boldsymbol{h}'_\lambda$，および $\boldsymbol{h}_l$ であり，出力は $P(\hat{y}=1)$ である．初めに，2 つの画像の差分の表現 $\boldsymbol{h}_{\text{diff}} = \text{CrossAttn}(\boldsymbol{h}'_\lambda, \boldsymbol{h}_\lambda)$ を得る．ただし，$\text{CrossAttn}(\cdot,\cdot)$ は以下のように定義される cross-attention を表す．

$$\text{CrossAttn}(\boldsymbol{X}_A, \boldsymbol{X}_B) = \text{softmax}\left(\frac{\boldsymbol{X}_A \boldsymbol{W}_q (\boldsymbol{X}_B \boldsymbol{W}_k)^\top}{\sqrt{d_k}}\right) \boldsymbol{X}_B \boldsymbol{W}_v \quad (1)$$

ただし，$\boldsymbol{W}_q$，$\boldsymbol{W}_k$，および $\boldsymbol{W}_v$ は学習可能な重み行列であり，d_k は $\boldsymbol{X}_B \boldsymbol{W}_k$ の次元を示す．次に，$\boldsymbol{h}_{\text{diff}}$ と $\boldsymbol{h}_l$

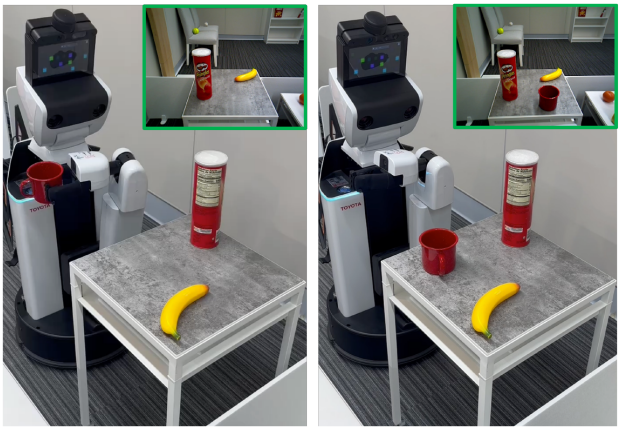


図 3 実験環境．左右の画像はそれぞれ物体操作の前後の状態を示す．“place a mug in front of the banana” などの指示文は物体操作後の状況を基に作成された．一人称視点画像の例はそれぞれの三人称視点画像の右上に表示されている．

間のアラインメント $\boldsymbol{h}_{\text{align}} = \text{CrossAttn}\left(\boldsymbol{h}_{\text{diff}}, \boldsymbol{h}_l\right)$ を得る．最後に，多層パーセプトロンを使用して $\boldsymbol{h}_{\text{align}}$ から $P(\hat{y} = 1)$ を算出し，本モジュールの出力とする．損失関数として交差エントロピー誤差を使用する．

5. 実験設定

実験では SP-RT-1 データセット [4] を利用した．ベースライン手法を UNITER-base/large [21]，InstructBLIP Vicuna-7B（InstructBLIP [3]），GPT-4 Turbo with Vision（GPT-4V [1]），Gemini 1.0 Pro Vision（Gemini [2]），および λ-Repformer [4] とした．InstructBLIP，Gemini，および GPT-4V については zero-shot で評価した．それぞれの手法は，以下の理由によりベースライン手法として使用した．UNITER は，Visual Question Answering タスクを含む多くの Vision & Language タスクにおいて優れた性能を示している．また，InstructBLIP，GPT-4V，および Gemini は大規模データセットを使って事前学習された代表的な MLLM であり，あらゆるタスクにおいて良好な結果が得られている．以上より，これらの手法を採用した．

包括的な評価のために zero-shot 転移設定における実世界環境でのモバイルマニピュレータを使った検証も行った．図 3 は WRS2020 [22] にて標準化されている環境を基にした実験環境を示している．また，トヨタの Human Support Robot を使用した．本実験において，すべての手法の検証は zero-shot で行われた．つまり，収集したデータを用いた追加学習は行わなかった．

6. 実験結果

6.1 定量的結果

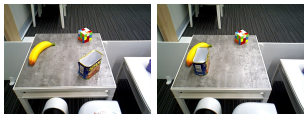
表 1 はベースライン手法と提案手法の比較の定量的結果を示す．表 1 より，SP-RT-1 データセットにおいて，提案手法の精度は 80.80%であった．また，UNITER-base，UNITER-large，InstructBLIP，GPT-4V，および Gemini の精度はそれぞれ 62.78%，63.52%，50.50%，63.90%，および 67.28%であった．このことから，提案手法は最も精度の高いベースライン手法の精度を 13.52 ポイント上回った．上記全ての結果において統計的に有意な性能差があった（$p < 0.001$）．

表 1 SP-RT-1 データセットにおけるベースラインと提案手法の定量的結果および zero-shot 転移実験の結果．最も精度が高いものを太字で示す．

手法	精度 [%]	
	SP-RT-1	Zero-Shot
UNITER-base [21]	62.78 ± 1.01	52 ± 1.6
UNITER-large [21]	63.52 ± 1.84	48 ± 1.8
InstructBLIP [3]	50.50 ± 0	50 ± 0
GPT-4V [1]	63.90 ± 1.04	59 ± 1.9
Gemini [2]	67.28 ± 0.80	53 ± 0.40
λ-Repformer [4]	74.50 ± 1.44	–
提案手法	**80.80** ± 0.86	**60** ± 1.8
Human performance	90	79

(i) “place water bottle upright”

ラベル：success



(ii) “move the rubik’s cube close to the banana”

ラベル：failure

図 4 Contrastive λ-Repformer の成功例．(i) は SP-RT-1 データセットに含まれる True Positive サンプルの例であり，(ii) は zero-shot 転移実験用データセットに含まれる True Negative サンプルの例である．各例において，左の画像は操作前のシーンを，右の画像は操作後のシーンをそれぞれ示している．

表 1 は zero-shot 転移実験での定量的結果も示している．これより，UNITER-base，UNITER-large，InstructBLIP，GPT-4V，Gemini，および Contrastive λ-Repformer の精度はそれぞれ 52%，48%，50%，59%，53%，および 60%であった．GPT-4V と Contrastive λ-Repformer は他手法よりも良い結果が得られた．さらに，Contrastive λ-Repformer の精度は GPT-4V をわずかに上回った．

本タスクを人間が行ったときの精度を検証するために，五人の被験者を対象に実験を行った．SP-RT-1 データセットでは，テスト集合から無作為に抽出した 100 サンプルに対し，被験者が SPOM タスクを行った場合の精度は 90%であった．zero-shot 転移実験に用いたデータセットでは，全サンプルにおいて人間による評価が行われ，精度は 79%であった．この結果から，SPOM タスクは人間にとっても難しいタスクであることが分かった．

6.2 定性的結果

図 4 は Contrastive λ-Repformer の成功例を示す．図 4 (i) は SP-RT-1 データセットに含まれる True Positive サンプルであり，図 4 (ii) は zero-shot 転移実験用データセットに含まれる True Negative サンプルである．図 4 (i) で与えられた指示は “place water bottle upright” である．このエピソードにおいて，マニピュレータは水筒を正しく操作し，直立になるように置いた．よっ

表 2 Ablation study の結果．最も精度が高いものを太字で示す．

モデル	Attention Mechanism	精度 [%]
(i)	Self-Attention	78.88 ± 1.05
(ii)	Cross-Attention	**80.80** ± 0.86

て，この例のラベルは success である．Contrastive λ-Repformer は，InstructBLIP を除くすべてのベースライン手法が success と予測できなかったこの例において，正しく success と予測した．図 4 (ii) に “move the rubik’s cube close to the banana” という指示文が与えられたサンプルを示す．このエピソードでは，マニピュレータがルービックキューブではなく青い缶を動かしたため，ラベルは failure である．このエピソードに対して Contrastive λ-Repformer は適切に予測したが，Gemini と InstructBLIP は誤って予測した．このエピソードより，提案手法は自然言語で書かれた文章と画像内の物体を適切にアラインできたことが示されている．

6.3 Ablation Study

Contrastive λ-Representation Decoder 内の cross-attention 操作の性能への寄与を調査するために ablation study を行った．この操作は，2 つの λ-Representation 間の差分の表現を作成するために使用した．表 2 にその結果を示す．

本実験では，cross-attention 操作の寄与を調査するために cross-attention 操作を self-attention 操作に変更した．表より，モデル (i) の精度は 78.88%であり，モデル (ii) よりも 1.92 ポイント低かった．これは，cross-attention 操作が画像間の差異を捉えることに適していることを示している．

7. おわりに

本研究では，物体操作前後の一人称視点画像と指示文が与えられた際に，open-vocabulary 物体操作の成功または失敗を予測するタスクを扱った．主な新規性として，2 つの画像間の違いを見つけ出し，その違いと指示文とのアラインメントを考慮することを可能にする Contrastive λ-Representation Decoder を導入した．さらに，提案手法の Contrastive λ-Repformer は，GPT-4V や Gemini などの代表的な MLLM を含むベースライン手法を上回った．将来研究として，この手法を様々な物体操作やナビゲーションタスクに適用することが考えられる．

謝辞

本研究の一部は，JSPS 科研費 23K03478，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参 考 文 献

[1] J. Achiam, S. Adler, S. Agarwal, et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2023.
[2] G. GeminiTeam, R. Anil, S. Borgeaud, Y. Wu, B. Alayrac, J. Yu, et al., “Gemini: A Family of Highly Capable Multimodal Models,” arXiv preprint arXiv:2312.11805, 2023.
[3] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, et al., “InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning,” NeurIPS, pp.49250–49267, 2023.
[4] 齋藤大地，神原元就，九曜克之，杉浦孔明，“マルチモーダル LLM および視覚言語基盤モデルに基づく大規模物体操作データセットにおけるタスク成功判定,” 第 38 回人工知能学会全国大会資料，pp.3O1OS16b02–3O1OS16b02，2024.
[5] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, et al., “VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models,” CoRL, pp.540–562, 2023.
[6] M. Shridhar, L. Manuelli, and D. Fox, “CLIPort: What and Where Pathways for Robotic Manipulation,” CoRL, pp.894–906, 2022.
[7] R. Korekata, et al., “Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks,” IROS, pp.3865–3872, 2023.
[8] F. Zeng, W. Gan, Y. Wang, N. Liu, and S. Yu, “Large Language Models for Robotics: A Survey,” arXiv preprint arXiv:2311.07226, 2023.
[9] K. Kawaharazuka, T. Matsushima, A. Gambardella, et al., “Real-World Robot Applications of Foundation Models: A Review,” arXiv preprint arXiv:2402.05741, 2024.
[10] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, et al., “RT-1: Robotics Transformer for Real-World Control at Scale,” arXiv preprint arXiv:2212.06817, 2022.
[11] Y. Qi, Q. Wu, P. Anderson, X. Wang, et al., “REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments,” CVPR, pp.9982–9991, 2020.
[12] Z. Liu, A. Bahety, and S. Song, “REFLECT: Summarizing Robot Experiences for Failure Explanation and Correction,” CoRL, pp.3468–3484, 2023.
[13] M. Shridhar, J. Thomason, D. Gordon, et al., “ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks,” CVPR, pp.10740–10749, 2020.
[14] C. Li, R. Zhang, J. Wong, et al., “BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation,” CoRL, pp.80–93, 2023.
[15] D. Driess, F. Xia, M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, et al., “PaLM-E: An Embodied Multimodal Language Model,” ICML, vol.202, pp.8469–8488, 2023.
[16] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, et al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances,” CoRL, pp.287–318, 2023.
[17] M. Shirasaka, et al., “Self-Recovery Prompting: Promptable General Purpose Service Robot System with Foundation Models and Self-Recovery,” ICRA, 2024.
[18] S. Haddadin, A. Luca, and A. Albu-Schäffer, “Robot Collisions: A Survey on Detection, Isolation, and Identification,” T-RO, vol.33, no.6, pp.1292–1312, 2017.
[19] A. Magassouba, K. Sugiura, A. Nakayama, T. Hirakawa, et al., “Predicting and attending to damaging collisions for placing everyday objects in photo-realistic simulations,” Advanced Robotics, vol.35, no.12, pp.787–799, 2021.
[20] M. Kambara and K. Sugiura, “Relational Future Captioning Model for Explaining Likely Collisions in Daily Tasks,” ICIP, pp.2601–2605, 2022.
[21] Y. Chen, L. Li, L. Yu, E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “UNITER: UNiversal Image-TExt Representation Learning,” ECCV, pp.104–120, 2020.
[22] “World Robot Summit 2020 Partner Robot Challenge Real Space Rules & Regulations,” 2020.