ポインタ局所探索による GUI Agent 向け 汎用 Grounding 後処理の検討

内藤美里[1] 宮本健[1] 三輪祥太郎[1]
[1]三菱電機株式会社
Naito.Misato@cb.MitsubishiElectric.co.jp
{Miyamoto.Ken, Miwa.Shotaro}@bc.MitsubishiElectric.co.jp

概要

LLM ベース GUI Agent では，動作計画から画面上座標への Grounding 精度が課題の 1 つである[1, 6]. 本研究では GUI での「ホバー時のハイライト」や「カーソル形状の変化」といった動的な反応を利用し，既存 GUI Agent に後付け可能な汎用 Grounding 後処理手法を提案する．Agent が出力した座標周辺でポインタをグリッド走査し，背景・カーソルの変化量から候補点を生成して，MLLM に最適候補を選択させる. 既存 GUI Agent である Agent S3 [4]に本手法を組み込み，PowerPoint [10]と FreeCAD [8]上のタスクで評価した結果，「長方形の右辺を選択せよ」のような精密 Grounding タスクにおいて成功例が増加し，単純なステップ数増加では代替できない効果が示唆された.

1 はじめに

GUI Agent は，自然文指示からクリックや入力操作を自動生成して汎用的なコンピュータ操作を実現する枠組みであり，その中で，自然文指示に対してどの画素をポインタ操作するかという GUI Grounding は重要な課題である[1, 6]．特に，CAD や図形編集ソフトで頻出する「図形の辺・頂点の選択」「細いスライダーハンドル操作」などでは，数十ピクセルの誤差が成功・失敗を分ける精密なクリック操作が求められる.

一般に，GUI は人間が操作しやすいよう設計されており，対象が選択可能な状態であることをユーザーに示すために，ホバー時のハイライトやカーソル形状の変化といった工夫が盛り込まれている．人間は, これらの視覚的フィードバックを利用しながら，目標とする UI 要素に向けてポインタを段階的に微調整していると考えられる.

本研究では，このような「人間向けに設計された GUI の動的な反応」を，既存の GUI Agent と補完的に利用する汎用 Grounding 後処理モジュールを提案する．人間が視覚的フィードバックを頼りに行っているポインタの微調整を自動で高精度に代替することを目指す．ベースの Grounding モデルやプロンプトを変更することなく，GUI Agent が推定した座標の近傍のみでポインタをグリッド走査し，背景およびカーソルの変化量に基づいて候補座標を生成し，各候補座標の局所スクリーンショットから MLLM にもっとも意図に合致する座標を選択させることで, 特に上述のような精密な GUI 操作における最終的な座標選択を補完する.

2 関連研究

GUI Grounding では, End-to-End での Grounding 推論や，テスト時スケーリング手法，CV モデルを用いた GUI 解析など多方面からの検討が行われてきた[1, 2, 6, 9]．ScreenSpot-Pro[6]は，その評価基盤として高解像度かつ専門アプリを含む多様なデータセットを提案し，既存モデルの限界を明らかにした.

RegionFocus [2]は，誤動作時に関心領域を複数のサブ画像へ分割し，各サブ画像に対する行動候補を画像上にマークとして可視化し, VLM に最適候補を選ばせる．さらに，RegionFocus の履歴もマークとして UI 画像上に残すことで，既探索領域を避けた多様な探索を促す．これは image-as-map mechanism と呼ばれ，VLM での効果的な選択を可能にする[2].

Iterative Narrowing [3]は, 推定座標を中心としたクロップを反復し，画像を徐々に縮小することで Grounding を精緻化する手法であり，追加学習なしに VLM の精度を向上させる[3]．SeeClick-Pro [6]は GUI の階層構造を考慮した視覚探索により性能を向上させた.

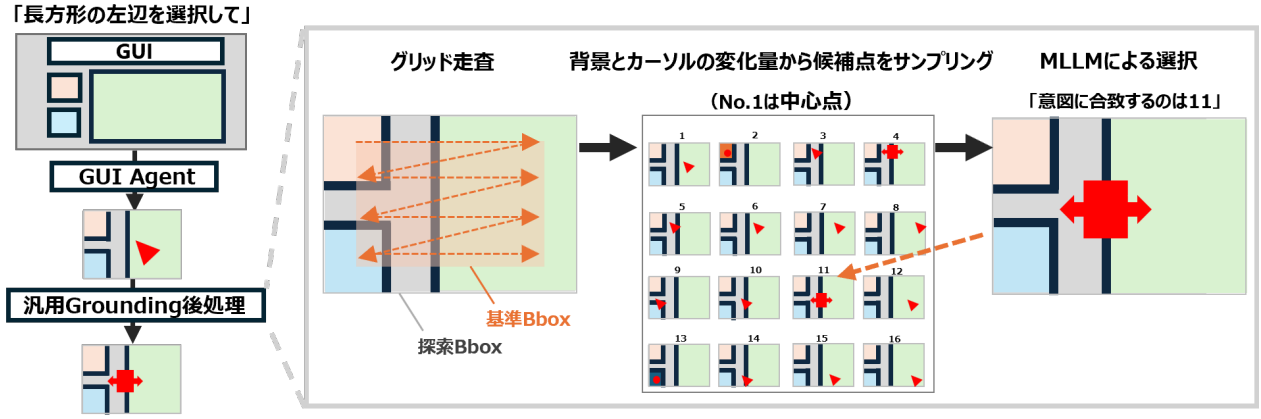

図 1　提案手法の概要図

一方，Agent S3 [4]は OSWorld [7]で SOTA の精度を示す GUI Agent フレームワークである．本研究は Agent S3 と，Grounding モデルとして UI-TARS-1.5-7B [5]を用いて，推論結果に対して汎用 Grounding 後処理モジュールを適用する．

3 提案手法

3.1 全体構成

提案手法の概要を図 1 に示す．GUI Agent が生成した GUI 操作指示の内，カーソル操作（移動・クリック・マウスダウン）について以下を行う．ベースの推論コードや LLM は変更しないため，既存 GUI Agent に後付け可能である．

1. GUI Agent が出力した座標の近傍でポインタをグリッド走査し，背景変化量マップとカーソル変化量マップを構成する．
2. 変化量マップから空間的に分散した候補座標を抽出し，各候補での局所スクリーンショットを取得する．
3. 中心点＋候補点のスクリーンショットを 1 枚のタイル画像にまとめる．
4. MLLM は、直前の計画テキストとタイル画像，対象のカーソル操作を示したメタデータを入力とし，もっとも意図に合致する番号を選択する．
5. 選択された番号に応じた座標で，実際のクリック等を実行する．

3.2 変化量マップと候補生成

推定された中心座標 (cx, cy) の周囲に基準 BBox を定義し，その外側にわずかに余裕を持たせた探索 BBox を差分計算用に用いる．基準 BBox をグリッド分割し，その格子点ごとに，探索 BBox の背景スクリーンショットとカーソルアイコンを取得する．基準状態との SSIM に基づく差分を計算することで，背景変化量マップとカーソル変化量マップを得る．これによりホバー時のボタンハイライトやカーソル形状の変化が数値として表現される．

変化量マップをサブブロックに分割したうえで，各値の連結成分ごとに重心を求め，それらの重心を候補点とする．同じグリッド位置に対応する候補点が複数存在する場合は 1 点に集約し，候補点集合を構成する．その上で，候補点同士ができるだけ散らばるように，すでに選択された点との最小ユークリッド距離が最大となる候補を貪欲に追加していき，最大 k 個の候補座標を得る．

3.3 タイル画像と MLLM プロンプト

各候補座標にポインタを移動させた状態で探索 BBox のスクリーンショットを取得し，タイル画像上に配置する．各セルには 1 からの連番ラベルを描画し，1 番は中心点に対応させる．

MLLM は，タイル画像と直前の計画テキスト，対象のカーソル操作を示したメタデータを入力として、「以前の推論意図とコードから，どの番号のスクリーンショットが元の意図に最も合うか」を選択する．選択された番号に応じた座標で，実際のクリック等を実行する．これにより，GUI Agent が推論した全体的な Grounding 結果を保ちつつ，最後の数十ピクセルを MLLM と GUI の動的な反応で補正する構成となる．

表 1　実験 1 精密タスクにおける「中心点＋候補点」のタイル画像に対する MLLM の推論結果

タスク名/SW 名	MLLM の選択理由	成功/失敗（Baseline は全て失敗）
"right edge" /PowerPoint	"Number 9 shows the cursor directly on the middle handle with the correct horizontal resize cursor, matching the intent to click the middle resize handle on the right edge."	成功
"upper right corner" /PowerPoint	"Screenshot 9 shows the resize cursor centered directly on the upper-right corner handle of the rectangle, matching the intended click on that specific resize handle."	失敗 : カーソルに変化があり，頂点とカーソルが重畳しているが，今回のタスクに適合するカーソルの変化ではない.
"right edge" /FreeCAD	"In image 6 the vertical right edge of the rectangle is highlighted blue, indicating the mouse is positioned directly on the edge segment between the corner points, matching the intended target.",	成功
"upper right corner" /FreeCAD	"Screenshot 4 has the cursor tip positioned directly on the red vertex point at the corner, best matching the intention to click the upper right corner vertex."	失敗 : カーソルの中心近くが頂点に重畳しており，先端は頂点を選択できていない.

4 実験・考察

4.1 実験設定

ベースの GUI Agent として Agent S3 [4]を用い，Grounding モデルには UI-TARS-1.5-7B [5]を設定した．MLLM は Azure OpenAI の gpt-5.1 を用い，`reasoning_effort="medium"`とした．その他設定は Agent S3 の公式ソースコードの基本設定に従う．提案手法における基準 BBox のサイズは画面サイズの 1/20 とし，その各辺を 32 ピクセルずつ拡張した領域を探索 BBox として用いた．グリッド分割は 10×20 で行い，サブブロック分割は 5×5 セル単位で行った．候補座標の最大個数 k は 15 とした．カーソル画像は Windows のシステム API を用いて取得した．

ScreenSpot-Pro データセット[6]では実行環境が用意されていないため，同ベンチマークを参考に，PowerPoint [10]と FreeCAD [8]上に 4 つの標準タスクを用意した．加えて，2 つの精密 Grounding タスクとして，描画された長方形に対するタスク"Select the right edge of the rectangle"，"Select the upper right corner of the rectangle"を自作した．画面サイズは 1920×1080 である．Grounding 成功は，スクリーンショットを目視し，人手で判定した．

表 2　実験 1 でのタスク成功数.

	標準タスク (全 4 タスク)		精密タスク (全 2 タスク)	
	Power Point	Free CAD	Power Point	Free CAD
Baseline	**3/4**	4/4	0/2	0/2
提案手法	2/4	4/4	**1/2**	**1/2**

4.2 実験 1：ステップ数 1 での比較

Agent S3 のステップ数（画面観測から実行までの行動サイクル数）を 1 に固定し，提案手法なし（Baseline）と、その出力に提案手法を適用した結果を比較した．その結果，タスク数は少ないものの，以下の差分が観測された（表 2）．

精密タスクにおける「中心点＋候補点」のタイル画像に対する MLLM の推論結果を表 1 に示す．PowerPoint・FreeCAD 共に"right edge"タスクではベースライン失敗・提案手法成功．"upper right corner"タスクでは両条件とも失敗したが，提案手法の候補集合の中には正解の点が含まれていた．このことから，精密な Grounding では，ポインタ近傍の探索によって正解の候補点を生成できる場合があること，および候補からの最終選択が MLLM の GUI 知識に依存していることが示唆された．

標準タスクでは 1 件，ベースラインが成功しているにもかかわらず、提案手法が微小なずれによって失敗した．ベースラインの出力では，カーソル画像は対象 UI に対して重畳していないが，カーソルの形状が適切に変化していた．一方，提案手法の出力では,カーソルが対象 UI に大きく重畳しているが，矢印カーソルの先端が対象 UI の左上に飛び出しており選択に失敗している（カーソルの形状に変化無し）．このことから，座標選択では「矢印カーソルの場合は先端座標が選択対象」，「○○を選択した場合，カーソルの形状が△△になる」等のカーソルに関する知識が重要であるが，MLLM がその知識を保持していなかったと考えられる．カーソルに関する知識を MLLM に事前に与えることで精度向上が見込まれる．

4.3 実験 2：ステップ数増加との比較

次に，PowerPoint の 2 つの精密タスクについて，提案手法を用いずに Agent S3 のステップ数上限を 5 まで増加させた．これは，ステップ数増加による試行錯誤で座標微調整が代替できるかを評価するためである． そのため，既定では Agent S3 で取得するスクリーンショットにカーソルは含まれないが，本実験用にマウスカーソルを重畳して描画した．

結果，本設定ではいずれのタスクも成功に至らなかった．"right edge"タスクでは，GUI Agent は長方形の右辺を選択できていないことは認識できているが，5 回のステップでは正確な座標を選択できずに終わった．"upper right corner"タスクでは，1 ステップ目の推論結果がわずかにずれていたが，正しく選択できたと判断して終了し，失敗している．

以上から，単純なステップ数増加は，提案手法が利用するような環境反応に基づく局所探索とは本質的に異なると考えられる．

5 まとめ

本研究では，GUI Agent のための汎用 Grounding 後処理として，GUI Agent が出力した座標近傍を実際に走査し，背景・カーソルの変化量に基づく候補点から MLLM に選択させる手法を検討した．Agent S3 [4]＋UI-TARS-1.5-7B [5]を用いた小規模実験から，特に長方形の辺をクリックするような精密 Grounding で，ベースラインが失敗するケースにおいて成功例が得られることを確認した． 一方で，候補集合に正解が含まれていても MLLM が選択しない事例や，わずかに悪化する事例も存在した．今後は，変化量マップの強度をより積極的に用いた候補スコアリングや，GUI 固有の知識をプロンプトに組み込むこと，RegionFocus [2]・SeeClick-Pro [6]との統合，さらに ScreenSpot-Pro 規模での大規模評価により，本手法の有効性を定量的に明らかにすることが課題である．

参考文献

[1] Chaoyun Zhang, Shilin He, Jiaxu Qian, Bowen Li, Liqun Li, Si Qin, Yu Kang, Minghua Ma, Guyue Liu, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, Qi Zhang. Large Language Model-Brained GUI Agents: A Survey. Transactions on Machine Learning Research, 2025. https://openreview.net/forum?id=xChvYjvXTp.
[2] Tiange Luo, Lajanugen Logeswaran, Justin Johnson, Honglak Lee. Visual Test-time Scaling for GUI Agent Grounding. IEEE/CVF International Conference on Computer Vision (ICCV 2025), 2025.
[3] Anthony Nguyen. Improved GUI Grounding via Iterative Narrowing, v7. arXiv : https://arxiv.org/abs/2411.13591, 2025.
[4] Gonzalo Gonzalez-Pumariega, Vincent Tu, Chih-Lun Lee, Jiachen Yang, Ang Li, Xin Eric Wang. The Unreasonable Effectiveness of Scaling Agents for Computer Use, v1. arXiv : https://arxiv.org/abs/2510.02250, 2025.
[5] ByteDance Seed. UI-TARS-1.5. (オンライン) 2025 年. https://seed-tars.com/1.5.
[6] Kaixin Li, Meng ziyang, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, Tat-Seng Chua. ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use. Workshop on Reasoning and Planning for Large Language Models, ICLR 2025, 2025. https://openreview.net/forum?id=XaKNDIAHas.
[7] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, Tao Yu. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024. ページ: 52040-52094.

https://proceedings.neurips.cc/paper_files/paper/2024/file/5d413e48f84dc61244b6be550f1cd8f5-Paper-Datasets_and_Benchmarks_Track.pdf.
[8] FreeCAD. (オンライン)
https://www.freecad.org/?lang=ja.
[9] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, Xiaojun Xiao, Kai Cai, Chuang Li, Yaowei Zheng, Chaolin Jin, Chen Li, Xiao Zhou, Minchao Wang, Haoli Chen, Zhaojian Li, Haihua Yang, Haifeng Liu, Feng Lin, Tao Peng, Xin Liu, Guang Shi. UI-TARS: Pioneering Automated GUI Interaction with Native Agents, v1. arXiv : https://arxiv.org/abs/2501.12326, 2025.
[10] Microsoft PowerPoint. (オンライン)
https://www.microsoft.com/en-us/microsoft-365/powerpoint.