Vol.2026-HCI-217 No.38
2026/3/11

Webページ上の要素がもつクリック可能性に対する人間による認知とVLMによる認識の一致度調査

徳原 眞彩[1,a)]　木下 裕一朗[1]　中村 聡史[1]

概要：企業や各種団体が Web サイトやアプリケーションなどを用いてサービス提供を行うことが一般的になってきたが，分かりづらいインタフェースも存在しており日々，人を困らせている．これまでの研究において，VLM が人のように使いにくいインタフェースを誤認することに着目し，Web サイトにおいてクリックが可能な要素の中で，ユーザがクリックできないと認知する可能性があるものを VLM を用いて推定し，Web ページのユーザへと強調表示により提示する手法を提案した．しかし，VLM の認識特性が人間の認知プロセスや振る舞いとどの程度合致しているか，あるいはどのように異なるかについては，これまで十分な検証がなされていなかった．そこで本研究では，Web ページ上のクリック要素に対する人間と VLM の認識結果を比較・分析することで，提案手法における VLM 利用の有効性およびその限界について議論する．分析の結果，VLM が検出した要素の多くは人間にとっても妥当であった一方，視覚的な手がかりが乏しい要素を見落とす傾向や，テキストの意味的文脈から過剰に要素を検出するといった，人間とは異なる認識特性を持つことが明らかになった．

1. はじめに

多くの企業や団体が，情報発信やサービス提供を目的として，Web サイトやアプリケーションなどを活用しており，それらの利用は一般的となってきた [1]．また，PC・スマートフォンの普及により Web サービスの利用者は増加し，利用者層の多様化や Web サービスの多機能化などが進んでおり，多くの企業は自社の商品やサービスを魅力的に見せるため，工夫を凝らしたデザインを取り入れることが多い．さらに様々なフレームワーク等が登場したことで，サイトのデザインはより複雑で凝ったものになっている．しかし，ユーザインタフェース（UI）に問題を抱えたサイトも多く，実際に利用する際に目的のページにたどり着くのに時間がかかったり，分かりやすい誘導ができていないためにユーザの意図していないページへ遷移させてしまったりすることがある．また，独自性の高いデザインはユーザにとって見慣れない UI となり，思った通りの動作を実行できない可能性もある．

サービスを提供する Web サイトにおいては，初めてそのサイトを訪問するユーザであっても瞬時に見慣れない UI を認知し，操作を考えること無く利用できることが望ましい．広川ら [2] は，直感的インタフェースの特徴としてユーザが操作対象を意識的な意味解釈をしなくても瞬時に認知し操作できることを挙げている．こうした UI を実現するため，UI 開発者が参照するために作られた UI ガイドライン [3][4] があるが，ガイドラインを参照することで分かりやすい UI の特徴を知ることができるものの，実際に開発した UI の大きさや配置，色合いに対してユーザがどのように認知するかを，UI の開発段階で予想するのは難しい．また，システムの開発者はその UI に慣れているため，ユーザ視点での分かりにくさに気づきづらいという問題がある．

図 1 と図 2 にユーザの利用の際に時間を要してしまう可能性のある UI の例を示す．図 1 に示すサイトは，上の 5 分野を示すイラスト群はクリック不可能であり，下の「ネットゼロ 5 分野の取組み」を押さなければ詳細が確認できない．この 5 分野のイラスト群（クリック不可）が，クリック可能性を示すシグニファイア [5] として機能してしまうため，無駄なクリック操作が何度も行われてしまう可能性があり，問題のある UI といえる．また，図 2 に示す例は，ユーザが購入申し込みを目的として利用するものであるが，青色で装飾されている「購入申込」は操作できない．ここでは「購入申込みはこちら」の部分を操作する必要があるが，操作可能性を示す矢印が文字列と重なっているため，操作可能であると判断される可能性が低くなっている．このような UI に対してデザイナにフィードバックを行う関

[1] 明治大学
Meiji University
[a)] mahirde.0905@gmail.com

© 2026 Information Processing Society of Japan

図 1 上 5 つの画像はクリックできず，ページ下部の「ネットゼロ 5 分野の取り組み」がクリック可能なサイト (inpex.com)

連研究は多く存在する [6][7]．しかし，既にリリースされているサービスに対してユーザが用いる UI 改善ツールについては十分に研究されていない．

著者らはこれまで，こうしたデザインによる不便さを解消するため，VLM（Vision-Language Model）が人のように，使いにくいインタフェースを誤認することに着目し，Web サイトにおいてクリックが可能な要素の中で，ユーザがクリックできないと認知する可能性があるものを VLM を用いて推定し，Web ページのユーザへと強調表示により提示する手法を提案してきた [8]．しかし，VLM の認識特性が人間の認知プロセスや振る舞いとどの程度合致しているか，あるいはどのように異なるかについては，これまで十分な検証がなされていなかった．そこで本研究では，Web ページ上のクリック要素に対する人間と VLM の認識結果を詳細に比較・分析し，VLM の認識特性が人間の直感的な認知とどの程度合致あるいは乖離しているかを明らかにすることで，提案手法における VLM 利用の有効性およびその限界について議論する．

2. 関連研究

2.1 UI が与える影響

UI が人々の認知や行動，評価に与える影響についての研究は様々ある．

Tichindelean ら [9] は，Web サイトのユーザビリティを分析した結果，情報の配置や色彩などのデザイン要素がユーザの視線行動と情報の記憶に大きく影響を与えたことを示した．また Riegler ら [10] は，モバイルアプリケーションの UI の複雑性を定量化し，UI 要素の数や密度，色の組み合わせといった視覚的要素がユーザの使いやすさや作業負荷に与える影響を調査した．その結果，UI の複雑さが高いほど作業効率が低下し，ユーザビリティが低く評価されることを明らかにした．Sundar ら [11] はクリックやスライド，ズームなどを含む 6 つの操作手法がユーザ体験に与える影響を分析し，スライドは記憶力を向上させ，ズームはユーザの評価が低い傾向を確認した．また，直感的で自然な操作手法がユーザ体験を向上させる一方，これらの手法を組み合わせた複雑な操作は評価を下げる可能性を示した．さらに，久保 [12] は「UX のハニカム構造」に含まれる，使いやすさや探しやすさが，EC サイトの利用

図 2 「購入申込」はクリックできず，「購入申込みはこちら」がクリック可能であるアプリケーション（「ナカペイ」スマートフォンアプリ）

においてブランド態度へ正の影響を与えることを明らかにした．

これらのことから，操作が直感的であると感じる要素には UI のデザイン性も含まれていると考えられるため，ユーザ体験の向上には UI が重要なことが分かる．また，UI はユーザ体験だけでなく，それらを扱うブランドへの印象も変化させる可能性があるため，ユーザが利用しやすい UI を提供することが重要である．本研究は，こうした Web サイトにおける UI の使いやすさに焦点を当て，ユーザの認知と操作の関係性について議論する．

2.2 UI の視覚的評価

作成した UI に対する視覚的な評価を得るシステムはリリース前にデザインの改善を行う際の手掛かりとなり，様々な手法が提案されている．

Duan ら [6] は VLM を用いて，UI デザインとガイドラインを入力することによって UI のフィードバックを行うシステムを構築し，微細なエラーの発見やテキスト改善に有用性があることを示した．また，Bisante ら [13] は Web インタフェースのユーザビリティ評価を支援するツール「CWGPT」を開発した．実験により，本ツールがユーザが利用しづらい UI を検出し，デザイナ初心者にとって有用であり，インタフェース設計スキルの向上に寄与する可能性があることを示した．Deka ら [14] はモバイルアプリのデザイン評価を効率化する「ZIPT」を提案しており，既存の Android アプリケーションに対して，ソースコードにアクセスすること無くユーザ操作データを収集し，視覚化することで，デザインのパフォーマンスや改善点を特定することを可能にした．Eldon ら [15] は，Web ページ上でユーザが特定の要素を見つけるまでの時間を高精度で予測するモデルを作成した．またモデルは画像データとターゲットの特徴を組み合わせた予測を行い，予測の中で生成される注目されやすい要素である「Saliency Heatmap」を

© 2026 Information Processing Society of Japan

提示し，デザインの改善に役立てることができることを明らかにした．

このように UI に対して視覚ベースの評価を行い，結果をデザイナにフィードバックすることによってデザインの改善を促す研究が行われているが，本研究では，利用しづらい UI をユーザに対してフィードバックすることを目的としている．

2.3 ユーザの行動予測

UI に対してユーザがどのような印象を持つのかを推測する研究も存在している．

Swearngin ら [7] はモバイルインタフェースにおいて，ユーザが要素をタップ可能と認識する確率を予測する深層学習モデルを開発し，その精度は 90.2%と高い評価を得ていることを示した．また，山中ら [16] は，Web ページの DOM 構造解析に基づき，モバイル端末上の UI 要素に対するタップ成功率を推定するツール「Tappy」を提案した．Wu ら [17] は「Never-ending UI Learner」を提案している．これは，システムがアプリケーションをクロールし，「タップ可能性」等を学習するモデルであり，既存の人手でアノテーションを行ったデータセットを越える精度を示し，長期間にわたるデザインの進化にも対応できる可能性を示した．Yuan ら [18] は Web ページ上での視覚的検索性能を予測するための深層学習モデルを開発した．このモデルは，画像データやターゲットの特徴を組み合わせ，ユーザが特定の要素を見つけるまでの時間を高精度で予測する．Zhou ら [19] はモバイルデバイス上のクリック行動を深層学習を用いてモデル化し，ユーザのクリック履歴や UI の構造情報などから次にクリックされる可能性の高い UI 要素を予測する手法を提案した．Bylinskii ら [20] は人間のクリックデータや重要性の評価から，デザイン要素の中でどの部分が視覚的に重要かを自動で判別できるモデルを開発した．

これらの研究は，実際のユーザの行動予測をデザイナにフィードバックすることで客観的なデータを利用したデザインの改善を支援するものである．本研究では，ユーザの行動予測を拡張機能に適応させることで，クリック可能予測が難しい要素をユーザに提示するような支援を目指す．

3. クリック要素の分類と本研究の目的

ユーザがクリック可能な位置をクリック可能であると認知できなければ，ユーザはその要素を見逃してしまい，目的のページにたどり着くことができなかったり，操作に時間を要してしまったりする．本研究では，サイト上のクリックに関する要素は表 1 のように分類できると考える．表 1 において，(1) と (4) はユーザの認知とシステムの挙動が一致しているため問題が無いが，(2) と (3) はユーザの認知とシステムの挙動のずれが存在するため，ユーザの利用を妨げる UI であると言える．具体例を挙げると，(2) は図 1 のようなものであり，(3) は図 2 のようなものである．ここで (3) は特に，一度クリックできないと思い込んでしまったうえ，他の位置に同じリンクを発見できなかった場合は，サイト全体から再度探し直さなければならないためより一層不便を強いる UI となってしまう．これまでの研究 [8] では，この (3) を検出してユーザに示す拡張機能を提案したが，本研究では，VLM の認識特性の検証に焦点を当て，以下のリサーチクエスチョンを設定する．

表 1 クリック要素分類

	クリック可能要素	クリック不可能要素
クリックできそうな要素	○ (1)	△ (2)
クリックできなさそうな要素	× (3)	○ (4)

RQ: クリックできそうだが機能しない要素（表 1 の (2)）や，逆にクリックできなさそうだが機能する要素（同 (3)）に対して，VLM は人間と同様の誤認傾向を示すか？

ユーザがクリックできると認知するかどうかについては機械学習など様々なアプローチがあるが，中村ら [21] は，VLM がユーザを惑わせる分かりにくいインタフェースに遭遇した時に，人と同じように間違ってしまう可能性を指摘している．実際に，図 2 について，VLM に「購入する際に優先的にクリックすべき箇所はどこか」というプロンプトで質問したところ，まず目立つ「購入申込」が提示され，次に「購入申込みはこちら」が挙げられるなど，人間が陥りやすい誤認と類似した挙動が見られた．そこで本研究では，この表 1 の分類に基づき，実際の Web サイト上の要素に対する VLM の判定結果と，人間の主観的な判定結果を比較することで，VLM が人間の視覚的認知の代替指標としてどの程度有効であるかを検証する．なお，本研究において「クリックできそう」とは，要素の形状や色といった視覚的特徴に加え，「ログイン」や「登録」といった行為を誘発する明示的な言語情報に基づいて，ユーザが直感的に操作可能であると判断する性質を指す．また，厳密には VLM の出力は計算機による推論結果であるが，本研究ではこれを人間の認知プロセスをシミュレートした「認知結果」とみなして比較を行う．

4. データセット構築

本研究では，人間と VLM の認識傾向を比較・分析するため，Web サイトのスクリーンショットにそれぞれがクリック可能であると判断した要素にアノテーションを行ったデータセットを構築した．

4.1 Web ページ画像の収集

多様なデザインが含まれる実世界の Web サイトを対象とするため，日本全国の各市区町村の公式ホームページか

© 2026 Information Processing Society of Japan

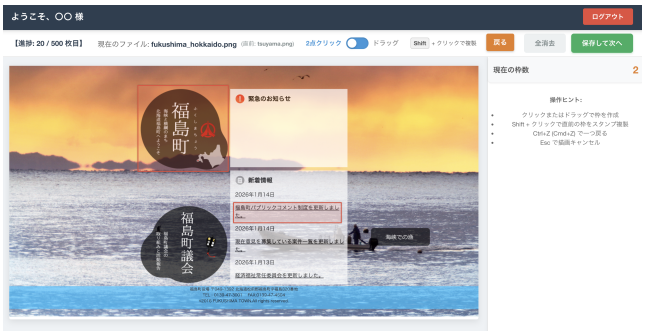

図 3 アノテーションシステム

ら様々な位置のスクリーンショットを収集した．市区町村の Web サイトは，自治体ごとにデザインのガイドラインや更新時期が異なるため，モダンなデザインから前時代的なデザインまで幅広く含まれており，本研究の調査対象として適していると考えられるためである．収集した画像は合計 1,000 枚であり，画像の解像度は 2880×1538 ピクセルである．

4.2 人間によるアノテーション

人間が Web ページ上のどの要素を「クリック可能（遷移可能）」と認識するかを調査するため，収集した画像に対してバウンディングボックスの付与を行った．アノテータは情報学を専攻する大学生および大学院生の計 4 名である．1 枚の画像に対して複数名の判断を反映させるため，4 名のアノテータがそれぞれ 500 枚ずつを担当し，全 1,000 枚の画像に対して各 2 名によるアノテーションを実施した．なお，アノテーション作業には図 3 に示すシステムを用いた．

タスクとして，画像内でクリックによりページ遷移が発生すると直感的に判断できる要素を全てバウンディングボックスで囲むよう指示した．なお，本研究では「ページ遷移」を伴う要素に焦点を当てるため，ポップアップ表示やタブ切り替え，画像の拡大表示など，URL の遷移を伴わない操作要素は対象外とした．

4.3 VLM によるアノテーションと位置情報の補正

VLM を用いたクリック可能要素の抽出にあたり，本分析では表 2 に示すプロンプトを用いた．このプロンプトでは，ボタンやリンクなどの「ページ遷移を伴う要素」を抽出対象とする一方，カルーセルやタブ切り替えなどの「画面内操作で完結する要素」を除外対象として記述している．

しかし，現時点での VLM が出力する座標情報は，要素の境界を画素単位で正確に捉えることが難しい．分析の際に物体検出の評価指標として用いる，領域の和集合に対する重なりの割合を示す IoU (Intersection over Union) や，正解領域に対する包含率を示す IoS (Intersection over Self) は，正解領域との重なり具合を厳密に評価する指標であるため，VLM が対象要素自体を正しく認識していても，座標のずれによって「不正解」と判定されてしまう場合がある．

本分析の主たる目的は，座標情報の生成精度を評価することではなく，VLM が Web ページ上のどの要素をリンクとして認識するかという認知的傾向を明らかにすることである．そこで本研究では，座標精度の問題によるノイズを排除するため，VLM が出力したテキストを参考に，著者が手動で正しいバウンディングボックスを付与し直したものを「VLM の認識結果」として採用した．

5. 結果

5.1 全体的な検出精度

本研究では，収集した 1,000 枚の Web ページ画像に対し，人間による判断結果と VLM による推定結果を比較した．なお，比較の基準として，2 名のアノテータが共にクリック可能と判断した要素のみを対象とする「AND 基準（Consensus）」と，少なくとも 1 名が判断した要素を対象とする「OR 基準（Union）」の 2 つを用いて評価を行った．

評価指標として，再現率（Recall）および適合率（Precision）を用いた．提案システムの目的は，プログラム的に取得可能な「実際のクリック可能要素」の中から，VLM が視覚的に認識できなかった（＝人間も見落とす可能性が高い）要素を検出し，ユーザに提示することである．この文脈において，各指標はシステムに対し以下の意味を持つ．

- **再現率 (Recall):** 人間が判断した要素のうち，VLM も同様に抽出できた割合．これが低い場合，VLM は「人間にとって明らかな要素」を見落としていることになる．その結果，システムはこれらを「難解な要素」と誤認してユーザへの提示を行ってしまい，ユーザへのノイズとなる恐れがある．

- **適合率 (Precision):** VLM が抽出した要素のうち，人間も同様に判断していた割合．これが低い場合，VLM は人間がクリックできないと判断する要素まで抽出してしまい，それが実際にクリックできる要素であった場合にユーザへの提示が行われなくなってしまう．

第 3 章で述べたように，クリックできなさそうだが機能する要素（表 1 の (3)）を見逃した場合，ユーザはサイト全体からの再探索を強いられるため，この手戻りを防ぐことは重要である．もし VLM が過剰に要素を検出してしまうと，システムはそれを「視覚的に自明である」と誤って判断し，本来提示すべきハイライト等を行わなくなってしまう．そのため，提案システムにおいては見逃しよりも過剰検出を考慮して，VLM の誤検出を最小限に抑えることに重きを置く必要があると考えられる．

表 3 に検出結果を示す．適合率は OR 基準において 0.977 となり，VLM のみが抽出した要素は 230 件であった．これは，VLM が「クリック可能」と判断した要素の多くは人

© 2026 Information Processing Society of Japan

表 2 クリック可能要素抽出のために VLM に入力したプロンプト

あなたは Web サイトのユーザビリティ調査を専門とするエンジニアです。提供されたスクリーンショットを解析し、ユーザが「別のページへ移動できる」と直感的に認識できる要素（リンクアフォーダンスを持つ要素）を特定してください。
このリストは後工程の作業者がアノテーション（枠線付け）を行うための指示書となります。「誰が見ても場所と対象が一つに特定できる」ように記述してください。

1. 抽出対象（リストに含めてください）
ページ遷移（別 URL への移動）を目的とした以下の要素を特定してください。
・ボタン状の見た目を持つ要素　・色付きや下線付きのテキストリンク
・ナビゲーションバーに含まれる各項目　・画像リンク（バナー、商品写真など）
※アイコンとその横のテキストが同じ遷移先を指す場合は、1 つの要素としてまとめて記述してください。

2. 除外対象（リストに含めないでください）
以下の UI 操作系要素のような、ページ遷移ではない要素は除外してください。
・ハンバーガーメニュー、カルーセル切り替え、ページトップへ戻るボタン
・タブ切り替え（「お知らせ」「イベント」などの表示切り替え）
・入力フォーム、検索窓、虫眼鏡アイコン　・言語切り替え、文字サイズ変更ボタン
※ただし、遷移か操作か判断に迷う場合は、抽出対象に含めてください。

3. 出力形式
前置きや挨拶は省略し、以下の例のような箇条書きリストのみを出力してください。ない場合は無しと出力してください。
・[ボタン] ヘッダー右上の「ログイン」と書かれた青いボタン
・[リンク] 本文中の「利用規約」という青文字のテキストリンク
・[画像] 中央にある「春のセール」と書かれた大きなバナー画像
・[ナビ] フッターにある「会社概要」のリンク

表 3 基準の違いによる検出結果の比較

項目	AND 基準	OR 基準
人間の判断数	10,884	11,954
VLM 検出数	10,081	
一致	9,436	9,851
人間のみ	1,448	2,053
VLM のみ	645	230
Recall (再現率)	0.867	0.828
Precision (適合率)	0.936	0.977

表 4 VLM による誤検出要素のカテゴリ内訳 ($N = 230$)

カテゴリ	件数	割合
文字 (Text)	121	52.6%
写真 (Photo)	56	24.3%
ロゴ (Logo)	29	12.6%
バナー (Banner)	13	5.7%
アイコン (Icon)	11	4.8%
合計	230	100.0%

間にとっても妥当であり，第 3 章で懸念された提示漏れのリスクを低く抑えられていると考えられる．一方，再現率は AND 基準でも 0.867 に留まり，約 13%の「人が 2 名ともクリック可能と判断した要素」が VLM によって抽出されていないことが明らかとなった．これは，現状の VLM では提案システムが視覚的に分かりやすい要素の一部まで提示対象として抽出してしまい，ユーザへの提示が余計に行われてしまう可能性が考えられる．

5.2 誤検出要素の詳細分析

VLM が過剰に検出した要素について，その視覚的な特徴を分析した．本分析では，2 名のアノテータが共にクリック可能と判断しなかった明確な誤検出に焦点を当てるため，OR 基準において誤検出と判定された要素の 230 件を対象とし，目視によるカテゴリ分類を行った．なお，誤検出の定義は，人がアノテーションを行った結果との IoU および IoS がいずれも 0.5 未満であるものとした．

分類結果を表 4 に示す．誤検出の中で最も高い割合を占めたのは「文字（Text）」であり，全体の過半数となる 52.6%（121 件）を占めた．また，「写真（Photo）」が 24.3%（56 件），「ロゴ（Logo）」が 12.6%（29 件）であった．この結果や目視での確認から，VLM の誤検出には以下の傾向があることが示唆された．

(1) **強調表現と文脈への過剰反応:** 「文字」カテゴリの多くは，見出しや注意書きなど，視覚的に強調されたテキストであった．図 4 に，VLM による誤検出例 (a) と，実際に機能する類似のボタン例 (b) の比較を示す．(a) の画像左下部分に位置する「重要なお知らせ」というラベルは，実際にはクリックできない単なる見出しである．しかし，(b) に示すような「緊急情報」ボタン（実際にクリック可能）と比較すると，「赤色の矩形背景」「白文字」「緊急性を訴える文言」といった共通点が多い．VLM は，(b) のような一般的なボタンの視覚的特徴や文脈を学習しているため，(a) のような類似した非リンク要素に対しても，一般化により「クリック可能」と判断してしまったと考えられる．一方，図 4(a) の要素は，今回はクリック可能な要素として実装されていないが，Web サイトの作成者がこの要素をクリック可能なリンクとして設定することは十分にあり得る．その場合，VLM が今回と同様に「クリッ

© 2026 Information Processing Society of Japan

(a) 誤検出されたラベル・クリック不可
（town.miharu.fukushima.jp）

(b) 類似する実際のボタン・クリック可能
（city.higashihiroshima.lg.jp）

図 4 VLM の判断根拠の分析．VLM は (b) のような実際に存在するボタンの特徴に基づき，(a) のような視覚的に類似した非リンク要素を誤認する傾向にある．

ク可能」と検出してしまうと，提案システムはこの要素を「視覚的にわかりやすい要素」と判断し，ユーザへの提示を行わない．もしユーザがこれを「単なる見出し」と認識していた場合，システムによる補助が得られないままリンクを見落とすことになる．このように，VLM が人間以上に「クリックできそう」と判断してしまう傾向は，システムが本来拾うべき「ユーザの気づきにくいリンク」を埋没させるリスクに繋がるため，問題となる．

(2) **非遷移画像の誤認:** 「写真」や「ロゴ」は，Web サイト上でリンクとして機能する場合も多いが，単なる装飾やヘッダー画像として配置されている場合も多い．VLM は「画像＝クリック可能」と判断してしまうケースがいくつか見られたが，人はクリック可能と判断していない要素がいくつか見られた．何も文字が書いていない画像や小さなロゴなどがこれらに該当するケースがあったが，実際にリンクである場合もあり，提案システムが理想的な動作をしない例となっていた．

表 5 VLM による見逃し要素のカテゴリ内訳 ($N = 1,448$)

カテゴリ	件数	割合
文字 (Text)	957	66.1%
アイコン (Icon)	241	16.6%
バナー (Banner)	204	14.1%
写真 (Photo)	31	2.1%
ロゴ (Logo)	14	1.0%
その他 (Other)	1	0.1%
合計	1,448	100.0%

5.3 未検出要素の詳細分析

人間がクリック可能と判断したにも関わらず，VLM が見落とした要素について分析を行った．本分析では，VLM の明確な見落としに焦点を当てるため，アノテータ間で判断が割れる要素を除外し，AND 基準において VLM が見落とした要素のうち 1,448 件を対象とし，同様に目視によるカテゴリ分類を行った．なお，未検出の定義は，誤検出と同様に人がアノテーションを行った結果との IoU および IoS がいずれも 0.5 未満であるものとした．

分類結果を表 5 に示す．未検出となった要素のうち，最も大きな割合を占めたのは「文字（Text）」であり，全体の 66.1%（957 件）に達した．また，「アイコン（Icon）」が 16.6%（241 件），「バナー（Banner）」が 14.1%（204 件）となった．

この結果や目視での確認から，VLM の未検出には以下の傾向があることが示唆された．

(1) **視覚的特徴の乏しいテキストリンク:** 「文字」カテゴリが圧倒的多数を占めた要因として，フッターのリンク集やリスト項目など，ボタンのような枠線や背景色をもたない「プレーンテキストに近いリンク」を VLM が認識できていないことが挙げられる．図 5 に見落としの例を示す．「新着情報」の下部に配置された各ニュース項目は，実際にはリンクであるが，下線や枠線といったクリックを誘発する視覚的手がかりが欠如している．人間は「新着情報の一覧」という文脈からこれらをリンクと認識できるが，VLM は見た目を重視するあまり，これらを見落とした可能性などが考えられる．また，下線や色の違いなどがあっても，見落とした事例がいくつか見られたため記事のタイトルなどの要素を VLM がクリック可能と判断するのは難しい可能性がある．

(2) **アイコン・バナーの認識漏れ:** 「アイコン」や「バナー」も一定数見落とされている．これらは，画像内のオブジェクト認識には成功していても，「それがクリック可能か」の判定において，単なる装飾画像と区別できていない可能性が示唆される．

© 2026 Information Processing Society of Japan

図 5 VLM による見落としの例．「ネイチャーポジティブ宣言」などのテキストリンクは，VLM に認識されなかった．シグニファイアの欠如などが原因だと考えられる．（town.kaneyama.yamagata.jp）

図 6 VLM の誤検出の例．「行政」「子育て」などのテキストを人間は見落としたが，VLM は文脈からリンクと判断した．この場合，システムはユーザへの提示を行わない．（town.miyota.nagano.jp）

5.4 文脈理解による過剰な推定

提案システムでは，VLM が検出できなかった要素を「人間にとっても発見困難な要素」とみなしてユーザに提示する仕組みを採用している．しかし，分析の結果，VLM が視覚的な見た目ではなく，テキストの意味的文脈に過度に依存して正解を導き出してしまうケースが確認された．

図 6 にその事例を示す．この Web ページでは，背景写真の上に「行政」「移住・引越し」といった白文字が配置されている．これらは実際にはクリック可能なリンクであるが，下線やボタンの枠線といった視覚的な手がかりが欠如しているため，アノテータは見落としていた．一方，VLM はこれらの要素をクリック可能として検出した．

これは，VLM が画像内の視覚的特徴だけでなく，読み取った「行政」や「子育て」といった単語が，自治体サイトにおいてナビゲーションメニューとして機能する確率が高いという知識を用いて推定した結果であると考えられる．その結果，提案システムにおいては「VLM が認識できている」と判定されるため，本来ユーザに提示すべき，見た目が分かりにくいリンクであるにも関わらず，提示が行われないという問題が起こる．

このことから，VLM を人間の視覚的認知のシミュレータとして利用する場合，VLM が持つ言語能力が，かえって純粋な見た目の評価を阻害する要因となり得ることが示唆された．

6. 考察

6.1 検出精度の傾向とシステムへの影響

分析の結果，アノテータのいずれか片方の判断を正解として扱う OR 基準において，適合率は 0.977 であった．第 3 章で述べた通り，提案システムは VLM が抽出できなかった要素を「見落としやすい要素」として提示する仕組みであり，VLM による過剰な検出は提示漏れに直結する．本分析における適合率は，VLM がクリック可能と判断した要素の多くが人間にとっても妥当であることを示しており，本来提示すべき要素が VLM によって「自明な要素」と判定され提示対象から除外されるリスクは，実利用ができる範囲に抑えられていると考えられる．このことから，VLM の認識結果をユーザへ提示すべきかどうかという判定基準として活用する手法の利用可能性が示唆された．

一方，再現率は AND 基準で 0.867 に留まった．これは，人にとって明らかな要素の一部を VLM が認識できていないことを意味し，システム上は自明な要素に対する過多な提示を引き起こす要因となる．しかし，第 3 章の分類における「クリックできなさそうだが機能する要素（表 1 の (3)）」を見逃すことによる手戻りの大きさを考慮すれば，実用上は提示漏れを最小化する観点から，現状の精度においても支援効果が見込めるものと考えられる．

6.2 視覚情報と文脈情報の乖離

事例分析から，VLM の認知プロセスと人間の判断との間には差異があることが確認された．金山町の事例（図 5）に見られるように，下線や枠線といったシグニファイアが欠如したテキストリンクを VLM が認識できなかった傾向は，VLM の判定が要素の見た目にやや依存している可能性があることを表している．

一方，御代田町の事例（図 6）では，人間が見落とした白文字のリンクを VLM が正確に検出した．これは，VLM が文字認識結果と自治体サイトという知識を照らし合わせ，周囲の構造からリンクであると推論する「意味的理解」において高い能力を有していることが示唆される．提案システムの設計上，VLM が意味的に正解を導き出してしまうことは，視覚的な分かりにくさを評価できなくなることを意味する．人も同じように考えクリック可能と判断することは可能であるが，今回の分析においては誤検出として扱われていた．VLM の文脈理解能力が，代表的な UI の特性を考慮せずに，提示漏れを生じさせる可能性があるという課題が本分析を通じて分かった．

6.3 VLM を用いた UI 評価における今後の課題

以上の傾向を踏まえると，VLM を人間の視覚的認知の

© 2026 Information Processing Society of Japan

代替指標として利用する場合，その知識レベルをいかに制御できるかが重要な課題となる．VLM は Web サイトの一般的な構造に関するバイアスを有しているため，デザインが不適切であっても文脈から正解を推論できてしまう．今後は，プロンプトによって VLM の役割を制限する手法や，文字情報と見た目の情報のバランスを調整するなどのアプローチにより，より人間の直感的な視覚認知に近い検出手法を構築していく必要がある．

7. まとめ

本研究では，Web ページ上のクリック可能要素に対する人間の認知と VLM による認識の一致度を調査し，VLM を分かりづらい UI の検出の指標として利用する際の有効性と課題を明らかにした．

データセットを用いた定量的な評価の結果，適合率は OR 基準において 0.977 であった．これは，VLM がクリック可能と判断した要素の多くが人間にとっても妥当であることを示しており，本来提示すべき要素が VLM によって自明な要素と判定され，提示が漏れるリスクは実利用ができる範囲に抑えられていることが示唆された．一方，アノテータ双方の判断が一致した要素を正解として扱う AND 基準において，再現率は 0.867 に留まり，人間にとって明らかな要素を VLM が見落とすことによる提示過多の懸念も確認された．

事例分析においては，VLM と人間の認知プロセスの差異が確認された．VLM は，下線や枠線といった視覚的なシグニファイアが欠如した要素を見落とす傾向にある一方，テキストの意味的文脈からリンクを正確に推論する高い能力を有していることが明らかになった．この過剰な文脈理解能力は，視覚的に分かりにくい要素を「理解可能」と判定してしまうという，提示漏れに繋がる課題を提示している．

今後の展望として，様々なアプローチによる精緻化が求められる．例えば，VLM に入力する画像を意図的に低解像度化やぼかし処理を施すことで，意味的な情報の読み取りを制限し，人間の視認性に近づける手法などが考えられる．

参考文献

[1] 相生公成：クラウド時代の IT 産業エコシステム，産業学会研究年報， Vol. 2019, No. 34, pp. 91–111 (2019).
[2] 広川美津雄 他：直感的インタフェースデザインの設計論の基礎的考察，日本感性工学会論文誌， Vol. 13, No. 5, pp. 543–554 (2014).
[3] Apple Inc.: Apple Design Tips - デザインヒント (2024). https://developer.apple.com/jp/design/tips/ Accessed: 2026-2-8.
[4] Google LLC: Google Design - デザインガイドラインとリソース (2024). https://design.google/ Accessed: 2026-2-8.
[5] Islam, M. N. et al.: Exploring the impact of interface signs' interpretation accuracy, design, and evaluation on web usability: A semiotics perspective, *Journal of Systems and Information Technology*, Vol. 16, No. 4, pp. 250–276 (2013).
[6] Duan, P. et al.: Generating Automatic Feedback on UI Mockups with Large Language Models, *Proc. of the CHI '24*, pp. 1–20 (2024).
[7] Swearngin, A. et al.: Modeling Mobile Interface Tappability Using Crowdsourcing and Deep Learning, *Proc. of the CHI '19*, pp. 1–11 (2019).
[8] 徳原眞彩 他：Web ページ上のクリック操作にまつわる BADUI の VLM を用いた改善手法，情報処理学会研究報告ヒューマンコンピュータインタラクション（HCI），Vol. 2025-HCI-212, No. 43, pp. 1–8 (2025).
[9] Tichindelean, M. et al.: A Comparative Eye Tracking Study of Usability—Towards Sustainable Web Design, *Sustainability*, Vol. 13, No. 18 (2021).
[10] Riegler, A. et al.: Measuring Visual User Interface Complexity of Mobile Applications With Metrics, *Interacting with Computers*, Vol. 30, No. 3, pp. 207–223 (2018).
[11] Sundar, S. S. et al.: User Experience of On-Screen Interaction Techniques: An Experimental Investigation of Clicking, Sliding, Zooming, Hovering, Dragging, and Flipping, *Human–Computer Interaction*, Vol. 29, No. 2, pp. 109–152 (2014).
[12] 久保麻子：EC サイト／アプリにおける UX がブランド態度に与える影響，マーケティングジャーナル， Vol. 39, No. 3, pp. 32–51 (2020).
[13] Bisante, A. et al.: Enhancing Interface Design with AI: An Exploratory Study on a ChatGPT-4-Based Tool for Cognitive Walkthrough Inspired Evaluations, *Proc. of the AVI '24* (2024).
[14] Deka, B. et al.: ZIPT: Zero-Integration Performance Testing of Mobile App Designs, *Proc. of the UIST '17*, pp. 727–736 (2017).
[15] Schoop, E. et al.: Predicting and Explaining Mobile UI Tappability with Vision Modeling and Saliency Analysis, *Proc. of the CHI '22* (2022).
[16] 山中祥太 他：スマートフォン用ウェブページとアプリにおけるタップ成功率推定ツール Tappy の実用化，インタラクション 2025 論文集，pp. 119–128 (2025).
[17] Wu, J. et al.: Never-ending Learning of User Interfaces, *Proc. of the UIST '23*, pp. 1–13 (2023).
[18] Yuan, A. et al.: Modeling Human Visual Search Performance on Realistic Webpages Using Analytical and Deep Learning Methods, *Proc. of the CHI '20*, pp. 1–12 (2020).
[19] Zhou, X. et al.: Large-Scale Modeling of Mobile User Click Behaviors Using Deep Learning, *Proc. of the RecSys '21*, pp. 473–483 (2021).
[20] Bylinskii, Z. et al.: Learning Visual Importance for Graphic Designs and Data Visualizations, *Proc. of the UIST '17*, pp. 57–69 (2017).
[21] 中村聡史：BADUI 診療所: カルテ 48 人なみに引っかかる AI さん，ヒューマンインタフェース学会誌， Vol. 26, No. 1, pp. 22–23 (2024).

© 2026 Information Processing Society of Japan