公園利用実態把握に向けた モデル規模の異なる VLM 生成説明文のクラウドソーシング評価

Crowdsourced Evaluation of Park-Usage Descriptions Generated by VLMs with Different Model Scales and Deployment Settings

寺岡 莉玖 † 細川 蓮 † 松田 裕貴 †‡ 安本 慶一 † 諏訪 博彦 †
Riku Teraoka Ren Hosokawa Yuki Matsuda Keiichi Yasumoto Hirohiko Suwa

1 はじめに

地域公園は，休憩，遊び，運動，交流，イベントなど，多様な活動が行われる公共空間である．公園の整備，維持管理，設備配置および住民への情報提供を検討する上では，利用者数のような定量情報に加えて，利用者がどの場所でどのように活動し，時間とともに利用状況がどのように変化したかを把握する必要がある．しかし，従来の目視調査は調査者の負担が大きく，長期間かつ複数地点を継続的に観測することは難しい．公園内の利用者数や活動状況を一定の基準に基づいて記録する体系的な観察手法も提案されているが，多数の時間帯や地点を対象とする場合には，依然として人的コストが課題となる [1].

カメラ映像と深層学習を用いることで，公共空間における人物の検出，活動認識，空間利用の可視化を自動化する研究が進められている [2, 3]．これらの方法は人数や位置を定量化する上で有効であるが，住民や管理者が映像を直接確認せずに状況を理解するためには,「数人が広場を通過した」「一部の利用者が同じ場所に滞在した」といった意味的な説明も有用となる．

近年，VLM は画像・動画と自然言語を統合し，映像内の対象，行動および時間変化を自然言語で生成可能となっている．[4, 5, 6]．一方，生成された説明文は流暢であっても，人数の誤り，存在しない人物・動物・物体の記述，映像から確認できない目的やイベント内容の推測を含む場合がある．また，より詳細な説明は状況把握を助ける可能性がある反面，具体的な記述が増えるほど誤情報を含む機会も増える．したがって，公園利用実態の説明では，一般的なベンチマーク指標だけでなく，人間がどのような説明を有用と判断し，どのような誤りを問題視するのかを人間中心の観点から評価する必要がある．

本研究では，公園利用実態把握への社会実装を見据えモデル規模および実行環境の異なる 3 種類の VLM を用い，同一の公園動画から日本語説明文を生成する．具体的には，ローカル環境で実行可能な小規模 VLM として Qwen3-VL-2B と Gemma 4 E2B，クラウド API を通じて利用する大規模モデルとして Gemini 2.5 Flash を利用した．続いて，モデル名を伏せた 3 つの説明文をクラウドソーシング参加者へ同時に提示し，動画の内容を最もよく説明している文章を 1 つ選択させるとともに，その理由を自由記述で収集する．対象映像は，日常時昼，日常時夜，雨天・曇天，イベント時の 4 条件であり，各条件の動画に対して 25 件ずつ，合計 100 件の回答を得た．

実験の結果，全体では Gemini 2.5 Flash による説明文が 51%で最も選択された．しかし，場面状況別に見ると，雨天・曇天では Qwen3-VL-2B，イベント時では Gemma 4 E2B がそれぞれ最も多く選択され，本実験で用いた 4 本の対象動画では，動画ごとに説明文の選択傾向が異なった．また自由記述では，人数，人物行動，背景および時間変化の正確性・具体性が肯定的に評価された一方，存在しない物，不正確な時刻・天候・設備の記述などが明確な減点理由として挙げられた．特にイベント時には，詳細な断定を避けた説明が，情報量は少なくても誤りが少ないという理由で選択される例が見られた．

本研究の主な貢献は以下の 3 点である．第一に，実公園の異なる環境条件を対象として，3 種類の VLM で生成された説明文を同時比較する人間選好評価を実施したことである．第二に，選択数だけでなく自由記述理由を収集し，公園利用実態の説明で重視される要素を整理したことである．第三に，本実験で用いた対象動画では，一貫して最も高く評価されるモデルは確認されず，動画ごとの選択傾向の違いを考慮した説明生成システムの設計が必要であることを示唆したことである．

2 関連研究

本章では，屋外公共空間における利用実態把握，VLM による動画理解と説明生成，クラウドソーシングを用いた説明文評価，および公共空間における AI の社会実装に関する研究を整理し，本研究の位置づけを明らかにする．

2.1 屋外公共空間における利用実態把握

公園利用の観察手法として，公園内の対象エリアを一定時刻に観察し，利用者数や活動状態，周辺状況を体系的に記録する SOPARC がある [1]．これは再現性のある

† 奈良先端科学技術大学院大学，Nara Institute of Science and Technology
‡ 岡山大学，Okayama University

観察を可能にする一方で，観測回数や対象地点が増えるほど人的負担が増大するという課題がある．

この課題に対し，Sun らは公共オープンスペースに設置されたカメラと深層学習を用いて，歩行者・自転車利用者の検出と活動位置の可視化を行っている [2]．さらに，スマートシティカメラを用いて社会インフラの利用を時空間的にマッピングする枠組みも提案されている [3]．また，群衆人数推定では密度マップを推定する CSRNet などが知られており，画像から人数を定量化する技術が蓄積されている [7]．しかし，これらの多くは人数，位置，行動クラスなどの構造化出力を中心としており，一般利用者が直感的に理解できる自然言語説明の品質評価を主な対象とはしていない．

2.2 VLM による動画理解と説明生成

Video-ChatGPT は，動画特徴と LLM を接続し，動画に関する詳細な対話・説明生成を可能にしている [4]．Video-LLaVA は，画像と動画を共通の特徴空間へ整列させることで，画像・動画理解を統一的に扱う枠組みを示している [5]．Video ReCap は，短いクリップから長時間動画までを階層的に処理し，異なる時間粒度の説明を再帰的に生成している [6]．これらの研究は動画説明能力の向上に寄与しているが，主に公開データセット上の自動指標やモデルベース評価を用いており，特定の社会環境で生成文がどのように受け止められるかは別途検証することが重要である．

動画キャプションの自動評価では，参照文との語彙一致だけでなく，映像と生成文の意味的一致を測る EMScore などが提案されている [8]．ただし，公園利用実態の説明では,「詳細だが一部に誤りがある文章」と「簡潔だが重大な誤りがない文章」のどちらが実用上好まれるかなど，利用者の目的に依存する判断が含まれる．このため，本研究では自動指標ではなく，複数候補からの人間選好と選択理由を分析する．

2.3 クラウドソーシングによる説明文評価

画像や映像の説明文に関する研究において，クラウドソーシングは説明文の収集だけでなく，生成された文章の品質評価にも利用されている．Aguirre らは，画像の内容を短時間で把握するための簡潔な説明文と，画像を詳細に理解するための包括的な説明文をクラウドワーカーから収集する方法を検討している [9]．具体的には，必要とする説明の詳細度を文章で指示する方法と，画像を 500 ミリ秒だけ提示した後に説明文を記述させる時間制約付きの方法を比較している．収集した説明文を別の評価者が正確性，流暢性および詳細度の観点から評価した結果，時間制約を設けることで，正確性と流暢性を大きく損なうことなく，短く要点を捉えた説明文を収集できることが示されている．この結果は，クラウドソーシングにおける作業指示や提示方法が，得られる説明文の情報量や内容に影響することを示している．

また，Kasai らは，画像説明文の人手評価を明確な基準に基づいて行う評価枠組み THumB を提案している [10]．同枠組みでは，説明文に含まれる情報が画像と一致しているかを表す正確性と，画像内の重要な情報をどの程度含んでいるかを表す網羅性を中心に，流暢性や簡潔性などを分けて評価する．これは，同じ説明文であっても，存在しない対象を記述しないことを重視する評価者と，より多くの重要な情報を含むことを重視する評価者とでは，評価結果が異なる可能性があるためである．さらに，詳細であるが一部に誤りを含む説明文と，簡潔であるが誤りの少ない説明文のどちらを高く評価するかによっても，選択結果が変化し得る．

このように，人手評価では，単に説明文の良さを尋ねるだけでなく，正確性，情報の網羅性，詳細性，簡潔性など，回答者がどの要素を重視したかを把握することが重要である．本研究では，同一動画について生成された 3 つの説明文を同時に提示し，最も良い説明文を選択させる相対評価を採用する．さらに，その選択理由を自由記述で収集することで，説明文の選好に影響した具体的な要因を分析する．

2.4 公共空間 AI の社会実装

公共空間のカメラデータを利用するシステムでは，精度だけでなく，プライバシー，説明責任，住民が結果へ異議を申し立てられる仕組みなども重要である．Alfrink らは公共 AI の社会実装において，システムを異議申し立てに開かれたものとする contestability の必要性を論じている [11]．公園映像の説明生成においても，個人を特定する情報や根拠のない属性推定を避け，観測可能な事実と不確実性を区別して提示する必要がある．本研究は，住民向け情報提示へ VLM を利用する前段階として，生成説明文に対する人間の選好と誤記述への反応を整理するものである．

3 評価実験

本実験では，公園利用実態の説明における VLM 生成文の評価傾向を明らかにするため，モデル規模および実行環境の異なる 3 種類の VLM が生成した説明文をクラウドソーシングによって比較した．同一の公園動画に対する 3 つの説明文から最良の文章を選択させ，その選択結果と自由記述理由を分析した．評価実験の流れを図 1 に示す．

3.1 比較モデル

比較に用いた VLM を表 1 に示す．本実験では，公園利用実態把握への社会実装を見据え，ローカル環境で実

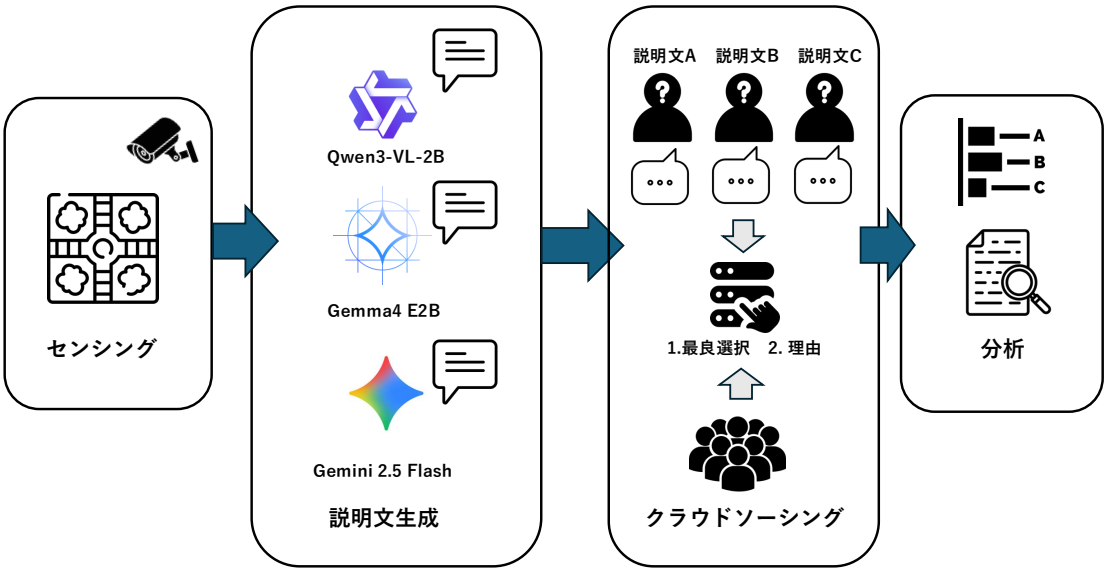

図 1: 評価実験

行可能な小規模 VLM として Qwen3-VL-2B と Gemma 4 E2B，クラウド API を通じて利用する大規模モデルとして Gemini 2.5 Flash を利用した．

Qwen3-VL-2B は，Qwen3-VL 系列に含まれる 20 億パラメータ規模の dense 型 VLM であり，テキスト，画像および動画を入力として扱う [12]．Gemini 2.5 Flash は，Google が提供するプロプライエタリなマルチモーダルモデルであり，テキスト，画像，音声および動画を入力できる．本実験では，クラウド API を通じて同モデルを利用した [13]．Gemma 4 E2B は，Google DeepMind が公開する open-weight のマルチモーダルモデルであり，端末上での効率的な実行を想定した Gemma 4 系列の小規模モデルである [14]．

なお，Gemini 2.5 Flash のパラメータ数は技術報告書では開示されておらず，Gemma 4 E2B の規模表記も一般的な総パラメータ数とは異なる．そのため，本実験はパラメータ数のみを統制した比較ではなく，モデル規模，アーキテクチャおよび実行環境の異なる VLM の生成説明文を比較するものと位置付ける．

3.2 対象とする公園動画

本実験で用いた動画は，既存の公開データセットから取得したものではなく，著者らが実際の公園に赴き，現地に撮影機器を設置して収集したものである．撮影対象は，大阪府豊能郡豊能町に位置する光風台中央公園（光風台二丁目公園）である．

対象公園は屋外環境であるため，時間帯や天候に伴う照度変化に加え，植栽，遊具および公園設備による遮蔽，利用者数や活動内容の時間的変動など，複数の要因が撮影映像に影響する．このような実環境では，人物が小さく映る場合や，遮蔽物によって一時的に見えなくなる場合があり，VLM による説明生成においても，人物数や行動，背景に関する誤認識が生じる可能性がある．そのため，本環境は，実際の公園利用実態を説明する上での VLM の有効性と課題を検証する対象として適している．

公園内には複数のカメラを設置し，現在も実際の公園利用を長期間にわたって撮影している．センサの設置位置を図 2 に示す．図中の番号は各カメラの識別番号を表している．本実験では遮蔽物が比較的少なく，撮影範囲全体の見通しが良いグラウンドを撮影するカメラ 8 の映像を使用した．すべての条件で同一視点のカメラ映像を使用することで，カメラの画角や設置位置の違いが，生成される説明文の評価に与える影響を抑えた．

なお，公園内には研究プロジェクト全体のデータ収集を目的として，カメラに加えてマイクおよび BLE スキャナも設置しているが，本実験ではカメラによって取得した動画のみを評価対象とする．また，撮影した動画の解像度は 800×600，フレームレートは 25 fps であり，1 本当たりの動画長は約 5 分である．

収集した動画の中から，撮影時刻，天候および公園の利用状況が異なる 4 本を選定した．各動画の条件と有効回答数を表 2 に示す．

具体的には，日常的な利用状況を撮影した昼間および夜間の動画に加え，雨天・曇天時の動画と，多数の人物や仮設設備が映るイベント時の動画を使用した．異なる

表 1: 比較した VLM

ラベル	モデル	規模の表記	提供形態
説明 A	Qwen3-VL-2B	2B	公開モデル／ローカル実行
説明 B	Gemini 2.5 Flash	非公開	クラウド API
説明 C	Gemma 4 E2B	E2B（実効規模表記）	公開モデル／ローカル実行

図 2: 光風台公園内に設置したカメラ等の設置位置と撮影方向.

映像条件を含めることで，各 VLM による生成説明文の評価傾向が，撮影環境や公園の利用状況によって変化するかを検証した．

3.3 説明文生成

各モデルには同一の公園動画を入力し，映像内で視覚的に確認できる情報に基づいて日本語の説明文を生成するよう指示した．

説明対象には，人の有無，おおよその人数，主な行動，人物の大まかな位置，滞在または通過の様子，人物の出入りや継続的な滞在といった時間変化を含めた．また，個人を特定できる情報を推測しないよう指示し，不明な情報は断定せず，出力は 2〜4 文程度の簡潔な日本語説明文とした．

実験に使用したプロンプトを以下に示す．

> あなたは屋外公共空間の動画を分析しています。
>
> 動画内で視覚的に確認できる内容のみを，2〜4 文で簡潔に説明してください。特に，人の有無，おおよその人数，主な行動，滞在/通過の様子，時間変化に注目してください。
>
> 映像から確認できない内容は推測せず，不明な場合は「不明」と書いてください。個人を特定しないでください。

3.4 クラウドソーシングによる評価

参加者は，評価対象動画を視聴した後，同じ動画について 3 モデルが生成した説明文を同時に提示した．モデル名が評価に影響しないよう，それぞれを説明 A，説明 B，説明 C として匿名化した．

参加者には，以下の 2 問へ回答を求めた．

1. 「この動画に対する説明文として，最も良いと思ったものを選んでください.」
   説明 A，説明 B，説明 C から 1 つを選択
2. 「なぜその説明文が良いと思ったのか簡単に教えてください．特に選ぶ上でどの点を重視したか (正確さ，詳しさ，読みやすさ等) がわかるように書いてください」
   自由記述

本実験では，各説明文を個別に数値評価する絶対評価ではなく，同一動画に対する 3 つの説明文から最も適切なものを選択する，相対評価を採用する．これにより，説明文間の差が小さい場合でも，回答者が総合的にどの説明を好むかを直接把握できる．また，選択理由を自由記述で収集することで，正確性，詳細性，簡潔性，読みやすさなど，選択時に重視された観点を調査する．

なお，動画を実際に視聴したことを確認するため，動画中に確認用画像を一時的に表示し，その画像の内容を回答させる注意確認問題を最後に設けた．本分析には，注意確認問題へ正しく回答した各動画 25 件，合計 100 件の回答を使用した．また，各参加者は 1 本の動画のみを評価し，同一参加者による複数条件への回答は含まれていない．

3.5 分析方法

各動画および全動画を対象として，モデル別の選択数と選択率を算出した．全体の選択数が 3 モデル間で均等であるかを確認するため，適合度のカイ二乗検定を行った．また，本実験で用いた対象動画と選択されたモデルとの関連を検証するため，4 本の対象動画と 3 種類のモデルによる分割表を作成し，独立性のカイ二乗検定を行った．

表 2: 評価対象動画

動画 ID	条件	主な評価上の特徴	有効回答数
V1	日常時・昼	人物の行動と時間変化が比較的確認しやすい	25
V2	日常時・夜	夜間照明下における人数および行動の説明	25
V3	雨天・曇天	視認性の低下，傘を差した人物および通過行動	25
V4	イベント時	多数の人物，テント等の構造物および活動の多様性	25

自由記述については，各動画における選択傾向を補足するために回答内容を精読し，繰り返し言及された評価理由や，生成文に含まれる具体的な誤記述を記述的に整理した．本稿では自由記述をカテゴリ別に符号化・集計する内容分析は行わず，選択結果を解釈するための探索的な分析として扱う．

3.6 倫理およびプライバシーへの配慮

本研究の評価対象は公園全体の利用状況であり，特定の個人を識別することを目的としていない．そのため，説明文生成時のプロンプトでは，個人を特定できる情報を出力しないことを指示している．

クラウドソーシング参加者には，回答が研究目的で使用されること，回答結果が匿名化された状態で分析されること，および参加条件を事前に提示している．

4 結果

各モデルが生成した説明文の例として，雨天・曇天時の動画に対する出力を表 3 に示す．本章では，Qwen3-VL-2B が生成した文章を説明 A，Gemini 2.5 Flash が生成した文章を説明 B，Gemma 4 E2B が生成した文章を説明 C としてクラウドソーシングの結果を示す．

4.1 説明文の選択結果

表 4 に，各動画条件における説明文の選択数と選択率を示す．全体では，説明 B（Gemini 2.5 Flash）が 51 件（51%）で最多であり，説明 A（Qwen3-VL-2B）が 26 件（26%），説明 C（Gemma 4 E2B）が 23 件（23%）であった．全体の選択数に対する適合度検定では，3 モデルが均等に選択されるという帰無仮説は棄却された（$\chi^2(2) = 14.18, p < .001$）．

また，条件別で見ると，4 条件すべてで同じモデルが優位だったわけではなく，日常時昼および日常時夜では説明 B，雨天･曇天では説明 A，イベント時では説明 C が最多となった．映像条件と選択モデルの独立性を検定した結果，両者には有意な関連が確認された（$\chi^2(6) = 36.24, p < .001$）．この結果から，本実験で用いた 4 本の動画では，最も多く選択される説明文が動画ごとに異なることが確認された．

4.2 選択理由の分析結果

自由記述にみられた動画別の主な選択理由を表 5 に示す．

日常時・昼では，説明 B（Gemini 2.5 Flash）について，「具体的かつ時系列的である」「動画全体の流れが分かる」「動画を見ていない人にも状況が伝わる」といった回答がみられた．人物の行動や出入りを時間順に記述している点が評価された一方，短く簡潔な説明で十分であるとして，説明 C を選択した回答も確認された．

日常時・夜でも，説明 B（Gemini 2.5 Flash）が人数を具体的に示し，人物の移動や人数変化を簡潔に記述している点が評価された．一方，他の説明文に含まれる「2 人の子ども」「夜中」「全体が暗い」などの表現について，映像と一致しないことを理由に説明 B を選択した回答もみられた．

雨天・曇天では，説明 A（Qwen3-VL-2B）が天候，傘を差した人物，人物の移動および時間変化を比較的正確に記述している点が評価された．一方，他の説明文に含まれる，映像には存在しない犬や，固定カメラであるにもかかわらずカメラが移動したとする記述が選択しなかった理由として挙げられた．

イベント時では，説明 C（Gemma 4 E2B）が多数の人物の歩行や滞在，活動量の変化など，場面全体を大局的に記述している点が評価された．他の説明文では，テントの色や位置，天候，イベント種別などに映像と一致しない，または根拠が不十分な記述があると指摘された．また，説明 C は具体性が低いものの，明確な誤りが少ないことを理由に選択した回答もみられた．

以上から，説明文の選択では，人物数，行動，背景および時間変化を具体的に記述することに加え，映像と一致しない対象や根拠の弱い推測を含まないことが重視されていたと考えられる．また，詳細な説明が常に選択されるとは限らず，対象や行動が多いイベント時の動画では，簡潔で明確な誤りが少ないことを理由に説明文を選択した回答もみられた．

表 3: 雨天・曇天時の動画に対して各 VLM が生成した説明文

ラベル	モデル	生成された説明文
説明 A	Qwen3-VL-2B	この動画は、曇りの天気の下、広い砂地の公園を捉えています。背景には住宅が並び、桜の木々が美しく咲いています。時間は午前 9 時 50 分頃で、空は曇っています。動画の途中、2 人の人物が歩く様子が見られます。一人は白い傘をさして歩いており、もう一人は黒い服を着て歩いているようです。二人は公園内をゆっくりと進んでおり、他の人物や動きは見られません。全体的に静かで穏やかな雰囲気です。
説明 B	Gemini 2.5 Flash	動画は日中の屋外広場を映しており、背景には住宅街と満開の桜の木が見えます。動画開始からしばらくは広場に人の姿はありません。中盤には、画面左側から犬を連れた人が現れ、広場の端を歩いた後、画面外へ退場します。終盤には、画面右下から傘をさした人が広場に入って中央に向かって歩き、同時に画面中央奥に別の一人が通過していくのが確認できます。
説明 C	Gemma 4 E2B	この動画は、広大な砂地の屋外公共空間を様々な角度から捉えたものです。映像には、特定の人物の活動や集団の動きは確認できず、主に砂地と背景の住宅地、そして春の桜が咲いている様子が映し出されています。時間経過に伴い、カメラの位置が移動しているように見えますが、空間自体に大きな変化は見られません。

表 4: 動画条件別の最良説明文の選択結果

動画条件	説明 A：Qwen3-VL-2B	説明 B：Gemini 2.5 Flash	説明 C：Gemma 4 E2B
日常時・昼	5（20%）	**18（72%）**	2（8%）
日常時・夜	3（12%）	**16（64%）**	6（24%）
雨天・曇天	**14（56%）**	10（40%）	1（4%）
イベント時	4（16%）	7（28%）	**14（56%）**
合計	26（26%）	**51（51%）**	23（23%）

表 5: 自由記述にみられた動画別の主な選択理由

条件	最多選択	主な肯定的評価	他文への主な指摘
日常時・昼	Gemini	人物の行動や出入りを時間順に記述し，動画全体の流れを把握しやすい	行動や時間変化に関する情報が少ない
日常時・夜	Gemini	人数，人物の移動および人数変化を具体的かつ簡潔に記述している	人数，年齢層，時刻および明るさに関する誤記述
雨天・曇天	Qwen	天候，傘を差した人物および通過行動を比較的正確に記述している	存在しない犬やカメラ移動に関する誤記述
イベント時	Gemma	多数の人物の移動や滞在を大局的に記述している	テント，天候，イベント種別に関する誤記述や根拠の弱い断定

5 考察

5.1 全体結果と対象動画による選択傾向の違い

Gemini 2.5 Flash は全体で最も多く選択され，日常時昼および日常時夜の動画で高い支持を得た．一方，雨天・曇天およびイベント時の動画では，他のモデルが生成した説明文が最も多く選択された．この結果から，特定の VLM がすべての動画において，一貫して高く評価されるとは限らないことが示唆された．

ただし，本実験では各条件につき 1 本の動画のみを評価しているため，天候や利用状況に応じて特定のモデルが優れると一般化することはできない．社会実装においては，単一モデルの生成結果をそのまま採用するのではなく，複数モデルによる生成結果の比較や，映像との整合性を確認する仕組みを導入することが重要だと考えられる．

5.2 モデル規模と説明品質の関係

本実験では，クラウド API 型の Gemini 2.5 Flash が全体で最も多く選ばれたが，小規模な Qwen3-VL-2B および Gemma 4 E2B も特定条件では最多となった．この結果は，モデル規模が大きい，または汎用ベンチマーク性能が高いことが，すべての公園映像で人間に好まれる説明を保証しないことを示唆する．

ただし，Gemini 2.5 Flash のパラメータ数は非公開であり，3 モデルはアーキテクチャ，学習データ，動画サンプリング方式，推論環境も異なる．したがって，本結果からパラメータ数の因果効果を直接結論づけることはできない．より厳密にモデル規模の影響を検証するには，同一モデル系列の 2B，4B，8B などを同じ入力処理・推

論条件で比較する必要がある．

5.3 詳細性と幻覚のトレードオフ

日常時昼では，具体的な行動と時間変化を含む説明が高く評価された．一方，雨天・曇天やイベント時では，存在しない犬，カメラ移動，誤った設備・天候，根拠の弱いイベント種別などの記述が大きく評価を下げた．このことから，説明生成では単純に情報量を増やすのではなく，視覚的根拠の強い情報だけを選択する必要がある．

住民向け説明では，断定的な誤情報は簡潔な情報不足よりも深刻に受け止められる可能性がある．そのため，モデルへ「不明な場合は不明とする」「映像から確認できない目的や属性を推測しない」と明示することに加え，人数や対象物に対する信頼度を出力させる，根拠となる時間区間を併記する，重要事実を別モデルで検証するといった幻覚抑制策が求められる．

5.4 公園利用実態把握への示唆

回答者が重視した人数，行動，背景，時間変化は，公園利用実態を住民や管理者へ伝える上でも重要な情報である．特に，人が一時的に通過したのか，一定時間滞在したのかを時間順に示す説明は，単一時点の人数だけでは得られない利用の意味を提供する．

一方で，公園利用状況の説明は，映像内のすべてを詳細に列挙する必要はない．住民向けの情報提示では,「現在は数人が広場を利用しており，大きな混雑はない」といった短い文章が適切な場合もある．したがって，今後は正確性だけでなく，利用者層と提示目的に応じた説明粒度を評価する必要がある．

5.5 社会実装上の課題

公共空間映像を用いた説明生成を社会実装するためには，モデル性能だけでなく，プライバシー保護や運用方法についても検討する必要がある．具体的には，個人を特定できる情報や根拠のない属性推定を出力しないこと，映像の利用目的，保存期間およびアクセス権限を明確にすることが求められる．また，生成文に誤りが含まれる可能性を利用者へ示し，必要に応じて訂正できる仕組みを設ける必要がある．

本研究において，小規模な公開モデルが一部の動画で高く評価されたことは，エッジ環境やオンプレミス環境での運用可能性を示唆する．一方，本研究では処理速度，消費電力，API 費用および通信遅延を評価していない．今後は，説明品質に加えて運用コストやデータ送信条件を含めた総合的な評価が必要である．

5.6 制約と今後の課題

本研究では，1 か所の公園に設置した同一カメラから，異なる条件の動画を 4 本選定して評価した．そのため，季節，撮影位置，人数分布および公園環境の多様性は限定されており，得られた傾向を一般化するには追加の検証が必要である．また，各モデル・各動画について限られた生成文のみを評価しているため，生成結果の確率的なばらつきは考慮できていない．

さらに，比較したモデルは規模だけでなく，アーキテクチャ，学習データおよび実行環境も異なるため，モデル規模のみの影響を分離することはできない．自由記述の分析も探索的な整理に留まっており，クラウドワーカーと実際の公園利用者や自治体職員とでは，望ましい説明の基準が異なる可能性がある．

今後は，対象動画と回答者数を増やすとともに，同一モデル系列における規模比較，複数回生成によるばらつきの評価，説明文の提示順のランダム化を行う．また，住民や自治体職員を対象とした評価を実施し，正確性，簡潔性，有用性および信頼性のうち，実運用で重視される要素を明らかにする．

6 おわりに

本研究では，公園利用実態の説明において人がどのような生成文を高く評価するかを明らかにするため，実公園で撮影した 4 条件の動画を対象に，3 種類の VLM が生成した説明文をクラウドソーシングにより比較評価した．合計 100 件の回答を分析した結果，全体では Gemini 2.5 Flash が最も多く選択された一方，対象動画によって選択傾向が異なることが確認された．

また，自由記述から，公園利用実態の説明では，人数，行動，背景および時間変化の正確性に加え，誤記述や根拠の弱い推測を含まないことが重視されることが示唆された．

今後は社会実装に向けて，映像条件に応じたモデル設計や出力設計の検討，住民や自治体職員を対象とした評価を実施する．

謝辞 本研究は，JST さきがけ（JPMJPR2465）および JST 共創の場形成支援プログラム (JPMJPF2115) の支援を受けたものである．

参考文献

[1] Thomas L. McKenzie, Deborah A. Cohen, Amber Sehgal, Stephanie Williamson, and Daniela Golinelli. System for Observing Play and Recreation in Communities (SOPARC): Reliability and Feasibility Measures. *Journal of Physical Activity and Health*, Vol. 3, No. s1, pp. S208–S222, 2006.

[2] Peng Sun, Rui Hou, and Jerome P. Lynch. Measuring the Utilization of Public Open Spaces by

Deep Learning: A Benchmark Study at the Detroit Riverfront. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2228–2237, 2020.

[3] Peng Sun, Gabriel Draughon, Rui Hou, and Jerome P. Lynch. Automated Human Use Mapping of Social Infrastructure by Deep Learning Methods Applied to Smart City Camera Systems. *Journal of Computing in Civil Engineering*, Vol. 36, No. 4, p. 04022011, 2022.

[4] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

[5] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

[6] Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video ReCap: Recursive Captioning of Hour-Long Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18198–18208, 2024.

[7] Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1091–1100, 2018.

[8] Yaya Shi, et al. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17929–17938, 2022.

[9] Carlos Aguirre, Amama Mahmood, and Chien-Ming Huang. Crowdsourcing Thumbnail Captions via Time-Constrained Methods. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI)*, pp. 36–48, 2022.

[10] Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. Transparent Human Evaluation for Image Captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3464–3478, 2022.

[11] Kars Alfrink, Ianus Keller, Neelke Doorn, and Gerd Kortuem. Contestable Camera Cars: A Speculative Design Exploration of Public AI That Is Open and Responsive to Dispute. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[12] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv*, Vol. 2511.21631, pp. 1–42, 2025.

[13] Gheorghe Comanici, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv*, Vol. 2507.06261, pp. 1–73, 2025.

[14] Gemma Team. Gemma 4 Technical Report. *arXiv*, Vol. 2607.02770, pp. 1–17, 2026.