EngagedSync：MR共有空間内ユーザの相互状況理解向上のための適応的な周辺環境情報抽出・転送システム

志村 魁哉[1,a)]　加納 一馬[1]　Tahera Hossain[1]　片山 晋[1]　浦野 健太[1]　米澤 拓郎[1]　河口 信夫[1,2]

概要：遠隔コラボレーションにおいて，複合現実（MR）技術の活用が進んでいる．MR 技術の大きな利点として，ユーザが現実空間を視認しながら遠隔ユーザとコラボレーションできることが挙げられる．例えば，会議中に部屋を歩き回りながら議論したり，飲食しながらの会話，手元の資料を確認しつつ意見交換を行うなど，現実空間での自然な行動を取り入れたコラボレーションが可能になる．しかし，MR コラボレーションには，現実空間の情報をどの程度共有するかという課題がある．現実空間やユーザの行動に関する情報を過剰に共有するとプライバシ侵害のリスクが生じ，逆に共有が不十分だと遠隔ユーザの状況を把握しづらくなり，コラボレーションの質が低下する恐れがある．この課題を解決するために，我々はユーザの周辺環境情報に基づいて周囲の物体を動的に抽出し，ユーザの映像と統合する MR コラボレーションシステムを提案する．プロトタイプの実装を通じた性能テストとユーザフィードバックにより，システムの有効性を確認する．

EngagedSync: Enhancing Mixed Reality Collaboration through Situation-Aware Object Extraction and Transfer

Kaiya Shimura[1,a)]　Kazuma Kano[1]　Tahera Hossain[1]　Shin Katayama[1]　Kenta Urano[1]
Takuro Yonezawa[1]　Nobuo Kawaguchi[1,2]

1. はじめに

遠隔コラボレーションシステムは，COVID-19 パンデミック以降，現代社会においてその重要性がさらに高まっている．働き方や生活スタイルの多様化が進む中で，物理的な場所に依存しないリモートワークが多くの企業や教育機関で導入され [1], [2]，オンラインでのコミュニケーションが標準的な手法として定着した．これにより，柔軟な働き方が可能となり，ワーク・ライフ・バランスの向上や多様な人材の活用が促進されるなどの効果が見られている．一方で，期待されたほどの効率性や生産性が得られないという課題もある．遠隔コミュニケーションは，対面コミュニケーションと比較すると非言語情報の欠如により意思疎通が困難になり，その結果，一部の組織ではオフィス勤務への回帰も見られている．こうした状況において，物理的距離を超えた効果的な遠隔コラボレーションの実現と，現実空間と仮想空間を統合する新たな手法の確立が求められている．

このような背景の中，現実空間と仮想空間をシームレスに統合する複合現実（MR: Mixed Reality）技術の活用が進んでいる [3], [4]．MR は，仮想現実（VR: Virtual Reality）と異なり，現実空間を視認しながら仮想空間での協働作業を可能にするため，現実空間の様子を確認しながら作業を進められる．この特性は，空間認識や精密な作業を伴うタスクにおいて特に有用であり，例えば部屋内の移動や部品の組み立てなどの作業において効果を発揮する．その結果，MR を活用したコラボレーションは創造性や作業効率の向上が期待されている．

MR コラボレーションにおける主な課題は，遠隔ユーザ

[1] 名古屋大学 大学院工学研究科
Graduate School of Engineering, Nagoya University
[2] 名古屋大学 未来社会創造機構
Institutes of Innovation for Future Society, Nagoya University
[a)] kaiya@ucl.nuee.nagoya-u.ac.jp

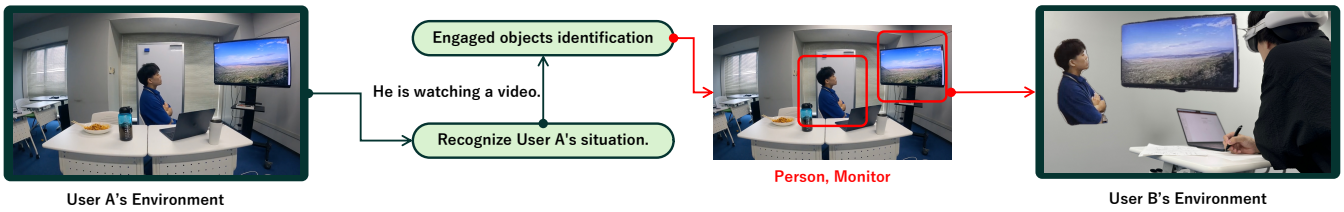

図 1 EngagedSync の概要：ユーザAが利用しているモノを自動認識し，ユーザAの本人の映像とともにセグメンテーションされ，ユーザBの視聴している MR 空間に重畳される．

の周囲の状況を認識して共有できていないため，相手の状況理解に齟齬が生じることである．現状の多くの MR コラボレーション手法は，ユーザのアバターを自然に遠隔空間に統合することに重点を置いている．しかし，遠隔ユーザの周辺にある物理オブジェクトとの相互作用は十分に考慮されておらず，これが遠隔ユーザの 周辺環境情報の認識の低下を招き，ユーザ間の同期的な行動を妨げる要因となっている．

これらの課題に対処するために，我々は EngagedSync という新たな MR コラボレーションシステムを提案する．このシステムでは，遠隔ユーザの行動に関連する物理オブジェクトを抽出し，その情報を遠隔ユーザ自身の映像と統合し，共有することで，遠隔ユーザの周辺環境情報の認識の向上を目指している．本アプローチにおける重要な特徴は，ユーザに関連するオブジェクトのみを選択的に転送する能力にある．これにより，ユーザ状況の理解が深まると同時に，プライバシ保護の実現を目指す．

本論文の貢献は，提案システムである EngagedSync の評価を通して，その有用性を明らかにする点にある．EngagedSync について，システム性能と印象評価の 2 つの側面から評価を行い，MR（複合現実）コラボレーションにおける有効な手法である可能性を示す．本論文の構成は次に示すとおりである．まず 2 章で関連研究について説明し，3 章で提案手法の説明をし，4 章で提案手法の評価を行う．最後に 5 章でまとめと今後の展望について述べる．

2. 関連研究

遠隔コラボレーションツールの需要は，グローバル化と技術革新に伴い増大しており，遠隔コラボレーション技術は依然として発展段階にあることが指摘されている [5], [6]．その一方で，企業や研究機関は国内外に拠点を持ち，異なる専門分野を持つ技術者が世界各地に分散している現状がある．このような現代社会において，対面でのコミュニケーションに近い遠隔コラボレーションシステムの開発が求められている．

2.1 物理オブジェクトの提示

遠隔コラボレーションにおいて，物理オブジェクトを遠隔ユーザに提示する場面が多く存在する．例えば，ビデオ会議においてオブジェクトの共有によって参加者間の一体感や協働の効率が向上することが示されている [7]．

先行研究では，物理オブジェクトを共有する研究が多くなされている [8], [9], [10], [11]．Hu らの ThingShare [12] は，ビデオ会議において無視されがちな物理オブジェクトを選択的に相手と共有する手法を提案している．また，Norris らの CamBlend [13] は広い視野角を持つカメラを利用し，注目している領域外にブラーエフェクトを施し，ユーザ間の周辺環境情報の認識を向上させている．しかし，これらの手法では，共有するオブジェクトをユーザが選択する手間が生じることや，それによってコラボレーションが一時的に中断されてしまうという課題がある．

2.2 共有空間の物理的不一致の解消

3D 点群や，アバターを使用したテレプレゼンス手法が提案されている．現実空間を再構築する共同 MR システムの開発も提案されており，ユーザ間のコプレゼンスを向上させている [4], [14]．しかし，これらの手法の問題の 1 つとして，ユーザが使用している環境の空間的配置がユーザによって異なるため，遠隔ユーザのアバターを重畳する際に，不自然な移動を引き起こす場合がある．これはジャメヴ効果とよばれ，コプレゼンスを下げることが指摘されている [15]．

この問題を解決するために，従来研究ではリダイレクテッドウォークをもちいて位置整合性を保つ方法が提案されている [16]．これは，ユーザの位置座標を変換することで，アバターを違和感なく重畳させるというものである．しかし，これらの方法は，あらかじめ似た配置の空間をセットアップする必要があるのに加え，複雑な配置の空間に対応できないという課題がある．また，これらの研究は，アバターの重畳手法に焦点を当てていて，ユーザの周辺環境情報が無視されてしまっているという課題も存在する．

3. EngagedSync

3.1 提案手法の概要

本研究の提案システムである EngagedSync は，共有するオブジェクトの動的な選択・抽出によって，MR コラボレーションにおける，遠隔ユーザの状況認識の向上と，プライバシの保護の両立を目的としている．以下は，EngagedSync

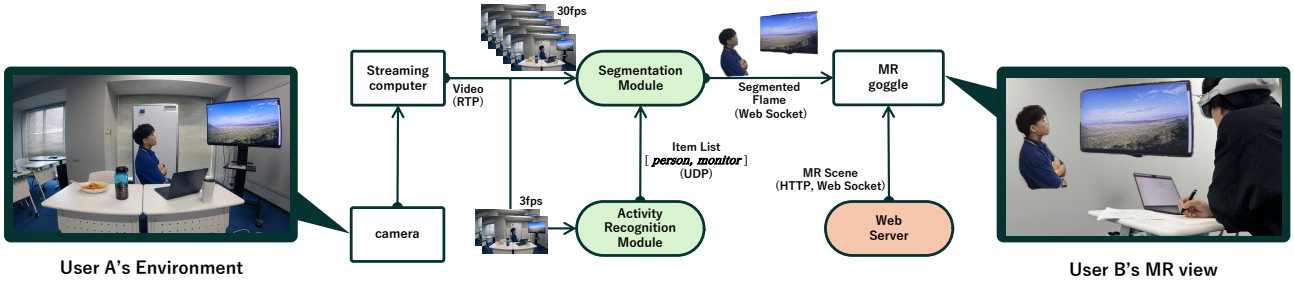

図 2 EngagedSync のシステム構成図

の主なシステムコンポーネントである．

- 配信環境
- アクティビティ認識モジュール
- セグメンテーションモジュール
- Web サーバ

EngagedSync のシステムアーキテクチャを**図 2** に示す．以降，説明のため，撮影される側のユーザをユーザ A，MR 空間でユーザ A の様子を観察するユーザをユーザ B とする．

まず，広角カメラを用いてユーザ A の環境を撮影する．撮影した映像は，アクティビティ認識モジュールとセグメンテーションモジュールの両方に送られる．アクティビティ認識モジュールでは，カメラ映像から一部のフレームを取り出し，ユーザ A の現在の活動に関係するオブジェクトを認識する．その後，“認識したオブジェクトを説明する短い文章” をリスト形式（本稿ではオブジェクトリストと呼ぶ）で出力する．セグメンテーションモジュールでは，オブジェクトリストに基づいて，カメラ映像の最新フレームに対してセグメンテーションを行い，セグメントしたフレームを配信する．ユーザ B は MR ゴーグルで MR 空間に参加し，ユーザ A の様子を確認する．次節以降では，各コンポーネントの設計と実装について詳しく述べる．

3.2 設計と実装

3.2.1 配信環境のセットアップ

本システムでは，ユーザ A の行動や周囲の環境をリアルタイムに捉えるために，広角カメラである GoPro を使用する．GoPro は配信処理を行うコンピュータに有線で接続されており，撮影した映像は後述するアクティビティ認識モジュールとセグメンテーションモジュールへ配信される．映像配信には GStreamer というオープンソースの配信用ソフトウェアを用いる．

3.2.2 アクティビティ認識モジュール

アクティビティ認識モジュールでは，受け取った動画から，ユーザ A の現在の行動を認識し，それに関連するオブジェクトを特定する処理を行う．行動認識には大規模視覚言語モデル（LVLM: Large Vision-Language Model）を使用する．本稿では，動画や画像の内容に関する質問への応答を生成できる Video-LLaVA [17] を採用した．

Video-LLaVa への入力は，配信用コンピュータからのカメラ映像とテキストプロンプトである．まず，カメラ映像について説明する．認識に使用する動画が短いと，ほぼ差分がない動画に対して推論を行なってしまうため．ユーザ A の活動を十分に認識できず，活動の認識精度が落ちてしまう．本稿では，配信フレーム数は毎秒 3 フレーム，Video-LLaVA 推論に用いるフレーム数は 8 枚に設定し，約 2.66 秒間の映像からユーザ A の活動を認識した．

次にテキストプロンプトについて説明する．Video-LLaVA に与えるプロンプトは”What is this person looking at?”である．このプロンプトにより，ユーザ A の活動に関連するオブジェクトを特定する．その後，Video-LLaVA が出力した文章を「, (カンマ)」および「and」で分割し，オブジェクトリストを得る．リストは UDP 通信でセグメンテーションモジュールに送信される．

3.2.3 セグメンテーションモジュール

セグメンテーションモジュールでは，現在の活動に関係するオブジェクトに対して選択的にセグメンテーションを行う．本稿では，多様なプロンプト形式に対応したマルチモーダルなセグメンテーションモデルである SEEM [18] を使用する．SEEM への入力は，配信用コンピュータから送られてくるカメラ映像の最新フレームとアクティビティ認識モジュールから送られてくる最新オブジェクトリストである．

オブジェクトリスト内の各要素に対して，テキストプロンプトによるインスタンスセグメンテーションを実行し，マスク画像を生成する．その後，各マスク画像に対して，ピクセル毎に OR 演算を適用し，一枚のマスク画像を合成する．また，テキストプロンプトを使用しないパノプティックセグメンテーションによって，フレーム内の全ての人物をセグメントする．セグメントされたフレームは WebSocket 通信でユーザ B の MR ゴーグルに配信される．なお，セグメンテーションモジュールの処理サイクルはアクティビティ認識モジュールがオブジェクトリストを生成するサイクルよりも短いため，同一のオブジェクトリストを複数フレームに渡って参照する．

3.2.4 Web サーバ

本システムでは，Web サーバとして ARENA [19] を使用

表 1 各動画に対するアンケート結果および適合率・再現率の加重平均

動画内でのユーザの活動	被験者が挙げた関連オブジェクト	挙げた被験者数	適合率の加重平均	再現率の加重平均
動画視聴をしている	monitor	13	1.00	0.68
	chair	5		
	laptop	1		
読書している	book	13	1.00	0.77
ホワイトボードに文字を書いている	whiteboard	13	0.93	0.61
	pen	12		
	eraser	9		
パスタを食べている	pasta	13	0.79	0.50
	chopsticks	12		
	dish	7		
	desk	1		
	bottle	2		
水分補給をしている	bottle	13	0.64	0.40
	pasta	1		
ノートパソコンを操作している	laptop	13	0.70	0.44
	desk	2		
	monitor	1		

する．ARENA は, マルチユーザ対応の XR アプリケーションの開発が容易な XR プラットフォームである．A-Frame という WebXR 用のフレームワークを使用して開発されており，MR シーンの編集が可能である．さらに，jitsi というビデオ会議ソフトウェアを組み込んでおり，XR 空間内でのビデオ通話が可能である．ARENA のシーン上に A-Frame のコンポーネントである plane オブジェクトを配置し，セグメントされたフレームを plane オブジェクトに貼り付ける．この時，画像のアルファ値を有効にすることで，マスクされた領域を透過させている．

クライアントの MR ゴーグルには Meta 社の Quest 3 を用いる．Quest 3 上のブラウザから Web サーバへアクセスして MR シーンを読み込むとともに，セグメンテーションサーバからセグメントされたフレームを受信して MR シーン上に描画する．ユーザ B の視界には，Quest 3 のパススルー機能によりユーザ B の現実空間が見えており，その空間上に，セグメンテーションサーバがセグメントしたユーザ A とオブジェクトが同時に重畳される．

4. 評価実験

本システムの性能を確認するため，システムの性能評価と印象評価の 2 つを実施した．本研究には，23〜52 歳の成人健常者 13 名（男性 11 名，女性 2 名）が参加した．

4.1 性能評価

4.1.1 実験内容

この実験では提案システムがユーザ A の行動に関連したオブジェクトを，動的に抽出できるかどうかを確認する．13 人の被験者に 6 つ動画を視聴してもらい，動画中の人の活動に関連するオブジェクトを列挙してもらった．動画は全て 5 秒間あり，1 つずつ動画をループ再生して被験者に視聴してもらった．被験者がオブジェクトを列挙し終えたら，次の動画へ進む形で実施した．その時，人によって表記に揺れのあるもの，例えば，モニター，ディスプレイ，テレビなどは，被験者に確認をとりながら同一の単語に統一するなどの処理を行なった．

4.1.2 評価結果

アンケート結果と評価結果を**表 1** に示す．被験者が挙げた物理オブジェクトを，Video-LLaVA に 100 回の推論を実行させ，適合率の加重平均と再現率の加重平均を算出した．この適合率の加重平均と再現率の加重平均は，被験者全体が挙げた物理オブジェクトを，その物理オブジェクトを挙げた被験者数に応じて重みを計算し，その重みを考慮して加重平均をとったものである．例えば，表 1 では，monitor, chair, laptop の重みはそれぞれ $\frac{13}{13}$, $\frac{5}{13}$, $\frac{1}{13}$，すなわち 1.0, 0.38, 0.08 となる．

各活動に対する適合率の加重平均と再現率の加重平均は表 1 の右側に示す通りである．動画視聴および読書では，適合率が 1.00 となり，検出された全てのインスタンスが正解であった．一方，再現率は動画視聴で 0.68，読書で 0.77 であり，一部のインスタンスが未検出であったことを示している．

ホワイトボードに文字を書く活動では，適合率が 0.93，再現率が 0.61 であった．適合率は高水準であるものの，再現率は中程度に留まった．これは，pen や eraser などの小さな物理オブジェクトが検出されなかったことが原因である．

食事，水分補給，ノートパソコン操作の活動では，適合率および再現率が他の活動に比べて低い値を示した．具体的には，食事の適合率は 0.79，再現率は 0.50 であった．水分

表 2　映像ごとの統計量の比較

比較映像	被験者への質問	Mean	Median	Std Dev	Min	Max
人だけの映像	リモートユーザの状況理解	2.77	3.00	1.05	1.00	4.00
	プライバシの安全性	3.92	4.00	0.92	2.00	5.00
	視界の良好さ	3.31	4.00	0.82	2.00	4.00
人と関連オブジェクトを合わせた映像（提案手法）	リモートユーザの状況理解	3.92	4.00	0.62	2.00	5.00
	プライバシの安全性	3.77	4.00	0.89	2.00	5.00
	視界の良好さ	2.38	2.00	0.74	2.00	4.00
すべての映像	リモートユーザの状況理解	4.69	5.00	0.46	4.00	5.00
	プライバシの安全性	1.62	2.00	0.62	1.00	3.00
	視界の良好さ	1.46	1.00	0.63	1.00	3.00

表 3　評価指標ごとの統計検定結果

Metric	Test Type	Comparison	Statistic	p-value
リモートユーザの状況理解	フリードマン検定	-	22.533	0.000013
	ウィルコクソン検定	人だけの映像 と 人と関連オブジェクトを合わせた映像	0.000	0.004017
	ウィルコクソン検定	人と関連オブジェクトを合わせた映像 と すべての映像	0.000	0.003892
	ウィルコクソン検定	人だけの映像 と すべての映像	0.000	0.000244
プライバシの安全性	フリードマン検定	-	21.800	0.000018
	ウィルコクソン検定	人だけの映像 と 人と関連オブジェクトを合わせた映像	2.500	0.317311
	ウィルコクソン検定	人と関連オブジェクトを合わせた映像 と すべての映像	0.000	0.001770
	ウィルコクソン検定	人だけの映像 と すべての映像	0.000	0.001689
視界の良好さ	フリードマン検定	-	15.571	0.000416
	ウィルコクソン検定	人だけの映像 と 人と関連オブジェクトを合わせた映像	0.000	0.013874
	ウィルコクソン検定	人と関連オブジェクトを合わせた映像 と すべての映像	5.000	0.008009
	ウィルコクソン検定	人だけの映像 と すべての映像	2.000	0.003269

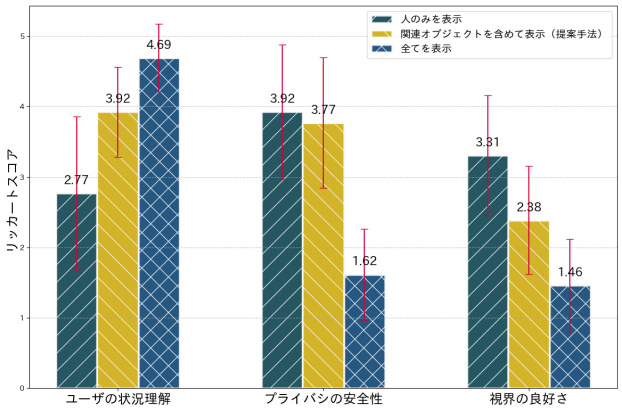

図 3　システムの印象評価結果

補給では，適合率が 0.64，再現率が 0.40 であった．ノートパソコン操作では，適合率が 0.70，再現率が 0.44 であった．

これらの結果から，関連オブジェクトを明確に見ている活動では高い適合率が得られたことが分かる．一方，活動が微細で多様性のある活動（水分補給，ノートパソコン操作）では，適合率および再現率が低下する傾向が見られた．特に，活動中，関連オブジェクトを見ない水分補給に関しては，適合率および再現率が低下する結果となった．

4.2　印象評価

4.2.1　実験内容

この実験では EngagedSync によるオブジェクト抽出の有効性を評価する．13 人の被験者には，MR ゴーグルを装着してもらい，1 つ動画の 3 つの映像パターンを MR 空間上で視聴してもらった．その後，印象に関する 3 つの質問に対して，5 段階のリッカート尺度で評価してもらった．

映像パターン

(1) ユーザ A のみを MR 空間に重畳した映像
(2) EngagedSync が抽出したオブジェクトとユーザ A を組み合わせて MR 空間に重畳した映像
(3) すべての映像を MR 空間に重畳した映像

質問項目

(1) リモートユーザの状況をどの程度理解できましたか？
(2) この人のプライバシを侵害する懸念をどの程度感じましたか？
(3) 視界が阻害されている感覚をどの程度感じましたか？

4.2.2 評価結果

評価結果を **図 3** に示す．また，この時の各統計量は **表 2** に示す．そして，検定結果を **表 3** に示す．リモートユーザの状況認識，プライバシの保護，および視界の良好さの 3 つの評価項目に関するデータを解析し，条件間の違いを検証した．各評価項目に対してフリードマン検定を実施し，有意差が認められた場合には，ウィルコクソンの符号付順位検定を用いて事後比較を行った．以下，各評価項目の結果について詳細に示す．

- リモートユーザの状況理解
  リモートユーザの状況理解のフリードマン検定の結果，統計量は 22.533，p 値は 0.000013 であり，条件間で有意な差が確認された（$p < 0.001$）．この結果を踏まえ，ウィルコクソンの符号付順位検定による事後比較を行ったところ，ユーザ A のみを重畳する場合とユーザ A と関連オブジェクトを合わせて重畳する場合との比較では，状況認識が統計的に有意に改善した（$p = 0.004017, p < 0.01$）．また，ユーザ A と関連オブジェクトを合わせて重畳する場合とすべての映像を重畳する場合との比較においても，状況認識が改善された（$p = 0.003892, p < 0.01$）
- プライバシの安全性
  プライバシの安全性に関するフリードマン検定の結果，統計量は 21.800，p 値は 0.000018 であり，条件間に有意な差が認められた（$p < 0.001$）．ウィルコクソンの符号付順位検定による事後比較では，ユーザ A のみの映像からユーザ A と関連オブジェクトを合わせて重畳する場合の比較では，両者に有意な差は見られず（$p = 0.317311$），この変化がプライバシや安全性の評価に明確な影響を及ぼしていないことが示された．
  一方で，ユーザ A と関連オブジェクトを合わせて重畳する場合とすべての映像を重畳する場合を比較したところ，プライバシと安全性の認識に有意な変化が見られ（$p = 0.001770, p < 0.01$），すべてのオブジェクトを共有することでプライバシの懸念が増大することが示された．
- 視界の良好さ
  視界の良好さに関するフリードマン検定の結果，統計量は 15.571，p 値は 0.000416 であり，条件間で有意差が確認された（$p < 0.001$）．ウィルコクソンの符号付順位検定による事後比較の結果，ユーザ A のみの映像のみを重畳する場合とユーザ A と関連オブジェクトを合わせて重畳する場合比較すると，後者では視界の妨害が大きくなることが示された（$p = 0.013874, p < 0.05$）．さらに，ユーザ A と関連オブジェクトを合わせて重畳する場合とすべての映像を重畳する場合を比較したところ，視界の妨害がさらに増大することが示された（$p = 0.008009, p < 0.01$）．

4.3 考察

本研究の結果から，EngagedSync は MR コミュニケーションにおいて，物理オブジェクトの動的な共有を通じて，相手の状況理解とコミュニケーションの質を向上させる可能性が示された．また，大規模言語モデル（LLM）の活用により，物理オブジェクトの抽出・共有プロセスを自動化・高度化できる可能性が示された．

4.3.1 相手の状況認識の向上とプライバシの保護のバランス

先行研究では，主にユーザの映像やアバターの重畳手法に焦点が当てられてきた．これらの手法は，遠隔地にいるユーザ間での存在感の共有や非言語的コミュニケーションを支援するものである．しかし，本研究は物理的なオブジェクトの共有に焦点を当て，新たな MR コラボレーションの可能性を提示した．

具体的には，EngagedSync は共有するオブジェクトを動的，かつ自動で選択できるため，ユーザ同士の議論を中断することなく，スムーズにコラボレーションを進めることができる．

4.3.2 制約

本研究にはいくつかの制約が存在する．現状，セグメンテーションの精度に関して，事前学習されたオブジェクトと未学習のオブジェクトでは検出精度に差が生じている．そのため，個別のチューニングが必要となる可能性がある．また，カメラの画角によっては，オブジェクトの正確な検出が困難になる場合もあり，撮影環境に依存する部分がある．モデルの誤検出は，プライバシの侵害につながる恐れがあるため，予防策の設計が必要である．さらに，システムの環境適応性やユーザインターフェースの改善，複数ユーザ間での協調を強化する機能の開発も課題として残されている．

5. まとめと今後の展望

本研究では，MR コミュニケーションにおいて，物理的なオブジェクトの共有を可能にする「EngagedSync」を開発し，その有効性を検証した．EngagedSync は，人の行動や状況に応じて，オブジェクトを動的に抽出・共有する機能を備えている．広角カメラとマルチモーダルな大規模言語モデル（LLM）を組み合わせることで，プライバシに配慮しつつ，相手の状況理解の向上を目指した．本システムの導入により，プライバシを保護しながら，相手の状況認識を効果的に向上させる可能性が示唆された．また，EngagedSync は，MR オブジェクトの重畳によるユーザの視界の負担を軽減し，より円滑なコミュニケーションを支援することが期待される．

今後の展望としては，360 度動画のストリーミングの実現が挙げられる．これを達成するためにはセグメンテーション技術が 360 度画像の歪みに対応できるように改良す

る必要や，比較的容量の大きな 360 度動画を効率的に処理する手法が求められる．また，オブジェクトの誤検出時にプライバシが侵害されないよう，レッドリストの導入や，動的検出と選択的共有のハイブリッドアプローチを検討する必要がある．さらに，ユーザインターフェースの改善により，システムの利便性と使いやすさを向上させることが求められる．複数ユーザ間での協調を強化するための機能の開発も，今後の研究の方向性として考えられる．これらの課題を解決することで，EngagedSync の実用性と応用範囲の拡大を目指す．

謝辞 本研究の一部は，JST CREST(JPMJCR22M4)，JST RISTEX(JPMJRS23K) および内閣府 SIP3 JPJ012495 に支援いただきました．

参考文献

[1] Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong, and Baining Guo. Virtualcube: An immersive 3d video communication system. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 28, No. 5, pp. 2146–2156, 2022.

[2] Audrey Labrie, Terrance Mok, Anthony Tang, Michelle Lui, Lora Oehlberg, and Lev Poretski. Toward videoconferencing tools for hands-on activities in online teaching. *Proc. ACM Hum.-Comput. Interact.*, Vol. 6, No. GROUP, January 2022.

[3] Barrett Ens, Joel Lanir, Anthony Tang, Scott Bateman, Gun Lee, Thammathip Piumsomboon, and Mark Billinghurst. Revisiting collaboration through mixed reality: The evolution of groupware. *Int. J. Hum.-Comput. Stud.*, Vol. 131, No. C, p. 81–98, November 2019.

[4] Andrew Irlitti, Mesut Latifoglu, Thuong Hoang, Brandon Victor Syiem, and Frank Vetere. Volumetric hybrid workspaces: Interactions with objects in remote and colocated telepresence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Gary M Olson and Judith S Olson. Distance matters. *Human–computer interaction*, Vol. 15, No. 2-3, pp. 139–178, 2000.

[6] Judith S Olson and Gary M Olson. How to make distance work work. *interactions*, Vol. 21, No. 2, pp. 28–35, 2014.

[7] Audrey Labrie, Terrance Mok, Anthony Tang, Michelle Lui, Lora Oehlberg, and Lev Poretski. Toward videoconferencing tools for hands-on activities in online teaching. *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6, No. GROUP, pp. 1–22, 2022.

[8] Christian Licoppe, Paul K Luff, Christian Heath, Hideaki Kuzuoka, Naomi Yamashita, and Sylvaine Tuncer. Showing objects: Holding and manipulating artefacts in video-mediated collaborative settings. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 5295–5306, 2017.

[9] Jens Emil Grønbæk, Mille Skovhus Knudsen, Kenton O'Hara, Peter Gall Krogh, Jo Vermeulen, and Marianne Graves Petersen. Proxemics beyond proximity: Designing for flexible social interaction through cross-device interaction. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[10] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. World-stabilized annotations and virtual scene navigation for remote collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, UIST '14, p. 449–459, New York, NY, USA, 2014. Association for Computing Machinery.

[11] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. In touch with the remote world: remote collaboration with augmented reality drawings and virtual navigation. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14, p. 197–205, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Erzhen Hu, Jens Emil Sloth Grønbæk, Wen Ying, Ruofei Du, and Seongkook Heo. Thingshare: Ad-hoc digital copies of physical objects for sharing things in video meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[13] James Norris, Holger Schnädelbach, and Guoping Qiu. Camblend: an object focused collaboration tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 627–636, 2012.

[14] Matthew Tait and Mark Billinghurst. The effect of view independence in a collaborative ar system. *Comput. Supported Coop. Work*, Vol. 24, No. 6, p. 563–589, dec 2015.

[15] Emily Wong, Jens Emil Sloth Grønbæk, and Eduardo Velloso. The jamais vu effect: Understanding the fragile illusion of co-presence in mixed reality. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, p. 2227–2246, New York, NY, USA, 2024. Association for Computing Machinery.

[16] Jens Emil Sloth Grønbæk, Ken Pfeuffer, Eduardo Velloso, Morten Astrup, Melanie Isabel Sønderkær Pedersen, Martin Kjær, Germán Leiva, and Hans Gellersen. Partially blended realities: Aligning dissimilar spaces for distributed mixed reality meetings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[17] 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, 2024.

[18] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once, 2023.

[19] Nuno Pereira, Anthony Rowe, Michael W Farb, Ivan Liang, Edward Lu, and Eric Riebling. Arena: The augmented reality edge networking architecture. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 479–488, 2021.