LooQuestion: LLM生成候補を用いたARグラス向けヘッドポインティング質問応答システム

土川 敦也[1,a)]　真鍋 宏幸[1,b)]

概要：AR グラスで周囲の情報を AI アシスタントに質問したいとき，音声入力や空中ジェスチャは公共空間で目立つ．本稿は，視点画像からマルチモーダル LLM が質問候補を生成し，ユーザは AR グラスに表示される質問選択肢をヘッドポインティングにて選択することで公共空間で目立つ音声入力や空中ジェスチャに依存しない AI アシスタントへの入力手法 LooQuestion を提案する．22 名が作成した視点画像に対する質問をクラスタ化してプロンプトを設計した．その結果，専用の追加学習なしでも，生成した質問選択肢は人手質問のクラスタを平均 55.0%カバーした．XREAL One/Eye/Beam Pro 上に Unity+FastAPI を用いて，8 件提示と追加生成を切り替えられるシステムの実装を行った．文字入力を排した質問選択型アプローチが，追加ハード不要で目立たない AR 質問応答を可能にすることを示す．

1. はじめに

大規模言語モデル（LLM）の発展により ChatGPT のようなサービスが普及し，検索エンジンと RAG を組み合わせることで幻覚を抑えつつ効率的な情報探索が可能になっている．Llama シリーズと Meta Quest を手掛ける Meta は，LLM ベースの AI アシスタントを統合した「Meta RayBan Display」を発表し，カメラ映像に対する音声質問が可能になっている．類似コンセプトの製品が続々と登場する中，将来的にはこのような AI 搭載 AR グラスがスマートフォンの機能の一部を置き換えると想定される．これらでは周囲に関する質問を，視点画像を参照しながら LLM がすることができるが，質問文の入力が不可欠である．タッチスクリーンを備えるスマートフォンと異なり，AR グラスは一般にタッチインタフェースを欠くため，音声入力や仮想キーボードへのジェスチャ入力が必要になる．しかし音声は発話のため目立ち，仮想キーボードのジェスチャも空中での大きな手の動きが必要で，公共空間での利用において操作者の心理的ハードルが大きい．商用製品にとどまらず，LLM と AR グラスを組み合わせてコンテキスト提示やロボット遠隔操作，プロアクティブなソーシャル支援を行う研究試作も増えているが，これらも音声入力に依存しており同様な課題を抱えている [1], [2], [3].

本研究では，AR グラスを装着するユーザの視点画像に含まれる情報を元にした質問群の生成を LLM が行い，その中からユーザに質問を選択させることで，ユーザによる文字入力そのものを不要にする．これにより，「ユーザが質問文を入力する」から「LLM 生成候補を選ぶ」へとパラダイムを転換し，公共空間でも目立たない質問応答を実現する．

2. 関連研究

本研究の関連研究は，画像から質問を生成する VQG（Visual Question Generation）と，目立たない入力に大別できる．従来の VQG は CNN やシーングラフの特徴を Seq2Seq ／ Transformer に渡す専用学習型で，ダブルヒント [4]，コントラスト学習 [5]，部分グラフ選択 [6] などで精度や多様性を高めてきた [7]．GPT-4V や Gemini などのファウンデーションモデルはゼロショット VQA で高性能かつ多言語対応を達成している [8].

目立たない入力の既存例には，歯クリック検出 [9]，無声発話センシング [10]，装着型タッチ面 [11] があり，いずれも追加ハードやキャリブレーションを要する．FlickPose はフリック＋手指ポーズで誤入力を減らすが動きは可視的である [12]．AI アシスタントへのユーザの入力にユーザの視線情報を統合することで，ユーザが注視対象を明示的に説明する必要がなくなることが報告されており [13]，視線入力は速度面でも優れるとされている [14]．一方で，視線入力には視線追跡用の追加ハードウェアが必要となり，コストが高くなりやすいという課題がある．そこで本研究は AR HMD の標準機能であるヘッドポインティングを採用

[1] 芝浦工業大学
[a)] ae19051@shibaura-it.ac.jp
[b)] manabehiroyuki@acm.org

し，水平方向に限定して社会的に目立たない操作とした．

3. 提案手法

AR 環境における AI アシスタントへの入力にて，音声入力や空中ジェスチャでは公共空間で目立つという課題に対し，ユーザが質問文を入力する代わりに LLM が質問候補を生成し，ユーザは他者から認識されにくい小さな動きを用いたヘッドポインティングにて選択を行う AR グラス向け AI アシスタントへの入力手法「LooQuestion」を提案する．従来の VQG のようにタスクごとのモデル学習を行わず，マルチモーダル LLM をプロンプト設計のみでゼロショット利用する点が特徴である．従来とは逆に「入力」ではなく「選択」を中心とすることで，(1) 音声入力や空中ジェスチャを伴う文字入力を排除して公共空間でも目立たない，(2) 質問を考える負荷を減らし即座に選択できるという利点を持つ．

システムは以下 2 つの主要システムから成る．第一に，カメラ付き AR グラスから送られた視点画像を取得し，質問生成プロンプトとともにマルチモーダル LLM へ送信して，予備調査で得たユーザ意図に沿った質問候補のバッチを得る VQG システム．第二に，カメラ付き AR グラスで視点画像を取得し，VQG システムに送り，VQG システムにより生成された質問候補を AR 視野に表示し，ユーザがヘッドポインティングで質問を選択する AR グラスシステムである．関連研究で述べた通り視線入力は速度と精度で優れるが専用ハードが必要である．一方ヘッドポインティングはほぼ全ての AR デバイスにて標準機能であり，自然な見回しに近い動きで目立たず，実装も簡便である．

4. 実装

実装としては大きく分けて，視点画像を元に質問選択肢を生成する VQG システムと，視点画像の収集・選択肢表示とヘッドポインティングでの選択を行う AR グラスシステムに分かれる．

4.1 VQG システム

VQG システムでは，視点画像を含むプロンプトを入力し，8 件の質問文を一括で生成する．ユーザの望む質問が含まれない場合は，異なる 8 件を再生成する．マルチモーダルモデルとして GPT-5 や Gemini 2.5 Pro や Claude 4.5 Sonnet を検討したが，MMMU ベンチマーク 84.2%などの性能および実際の質問生成精度から GPT-5 を採用した．

4.1.1 予備調査

効果的なプロンプト設計のため，AR グラス装着時に実空間の物体や人物に対しどのような質問を思いつくかを調べる予備実験を行った．VQG では VQG-COCO[15] のように第三者視点の一般画像質問が多いが，本システムには一人称視点の質問パターンが必要である．そのため 20 代の 22 名に AR グラス着用を想定させ，日常品から重機，人物を含む 15 枚の視点画像を提示し，各画像について 2–3 件の自由形式質問を収集した．得られた質問は OpenAI text-embedding-3-small で文章ベクトルを作成，文章ベクトル間のコサイン類似度 0.8 以上を同義としてクラスタリングし，画像ごとに上位約 4 クラスタの代表質問を抽出した．図 1 に小売店のプリンタ陳列画像で得られた代表クラスタを示す．

図 1 小売店プリンタと，そのプリンタへの人間による質問のクラスタの例

4.1.2 プロンプト設計

プロンプトには，状況説明と質問生成の意図を明記し，モデルの事前知識を引き出す．また，予備調査データのリークを避けるため，例示（one-shot）は予備調査とは異なるテーマや質問文を用い，返答スタイルの固定のみに利用した．これらを踏まえ，頻出する 4 種類の質問タイプが出力されるように調整した．．調整後，各視点画像につき 8 件生成した質問が 4 つの頻出質問タイプを内包する割合は最大 100%，最小 25%，平均 55.0%となった．同時生成数を 4/8/12/16 で比較すると，4 件で 45%，8 件で 55.0%，12 件と 16 件で各 51.7%となり，8 件が最もカバレッジが高かった．（図 2）12 件や 16 件では方向性が分散し，人間による質問のクラスタと類似の質問を網羅できなかった．

図 3 に最終的にカバレッジが最も高くなり，実装に利用したプロンプトを示す．

AR グラスから送られてきた画像をプロンプトに埋め込み OpenAI API の GPT-5(reasoning effort: medium) で推論し，質問候補のリストを JSON で AR グラスに返却する WebAPI を Python の FastAPI を用いて構築した．

4.2 AR グラスシステム

カメラ付き AR グラスとして XREAL One, XREAL Eye, XREAL Beam Pro を用いた．XREAL Eye は XREAL One に接続するカメラモジュールであり，XREAL Beam Pro は Android ベースで XREAL One と組み合わせ AR アプリケーションを実行する．AR アプリケーションは

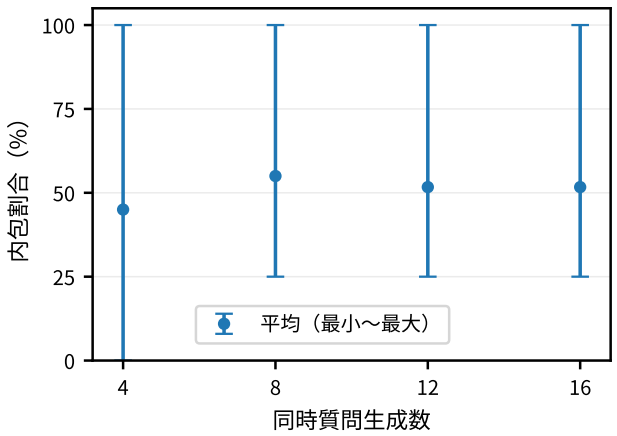

図 2　同時質問生成数と内包割合

SYSTEM PROMPT

あなたはカメラ搭載スマートグラスに搭載されたAIアシスタントです。
結果は必ず次のJSON形式で返してください:
{
"questions": [
{ "question": "...", "focus": "..." }
]
}
focus には質問が注目しているポイントを短く記述してください。

USER PROMPT

カメラで撮影した画像に映っている要素、特に画像中央にあるものを観察し、ユーザーが自然に抱きそうな疑問を日本語で{count}個作成してください。
ユーザーは何か明確な疑問を抱えている状態でカメラで画像を撮っている。撮影自体への質問はしないようにしてください。
質問は重複を避け、最も一般的に質問するであろう順に、敬体を使わずに簡潔にまとめてください。
ただし、あくまでユーザーは質問したい意図があり、その視点で画像を撮っているので、「ここはどこ」などのユーザーが明らかに知っている情報は質問しないようにしてください。

例)
入力：机の上にリンゴがある画像
出力：
- リンゴの品種は？
- 1個何円？
- リンゴを使った美味しいお菓子は？

注意事項：
- 基本的には素朴な質問を心がけてください。
- 一般常識で分かるものでないものの場合には、これは何？などの質問も検討してください。
- ただし、要求される個数が多ければ、最初の4個以降は、より発散した質問を後半では、考えて出力するようにしてください。
- 発散には画像中央のもの以外に対しても考えることを含みます。

USER PROMPT

図 3　VQG システムプロンプト

Unity で実装した．XREAL Eye を用いれば 6DoF も可能だが，本研究では XREAL Eye は視点画像取得のみに使用し 3DoF アプリとして実装した．視点画像取得は VQG システム側のサーバーからの信号によりトリガーされる．

不自然な垂直方向の動きを避けるために，質問候補は水平に視野内に配置することで，目立たなさを維持しながら AR グラスで質問を選択できるようにした．誤入力防止のため，候補上で 1 秒以上の頭部向きの固定を要求し，意図しない頭揺れでは選択が確定しないようにした．意図する質問がない場合は，追加生成の選択肢を選び，新たな 8 件を生成する．

4.3　実機動作

XREAL One/Eye/Beam Pro 上で机上の小物を対象に実機動作を確認した．Unity アプリは受信後に候補を視野中央付近へ水平配置して提示し，1 秒間の頭部向きの固定で選択を確定する（図 4）．

VQG システムのサーバーのダッシュボード画面より撮影指示を送り，カメラシャッターから候補一覧が視野に現れるまでのエンドツーエンド遅延を簡易計測したところ，平均で約 38.2 秒であった．

図 4　ユーザ視点での選択肢 UI（視点画像と実表示 UI を合成）

5.　議論

提案手法は，視点画像に基づいて LLM が質問候補を提示し，ユーザが水平方向のヘッドポインティングで選択するという「選択中心」の AR インタラクションを実現する．一方で現状のプロンプト設計では，人物や玩具など解釈の幅が広い対象に対して質問の観点（属性・用途・関係性など）が分散し，ユーザが意図する質問が候補に含まれにくい場合がある．したがって今後は，ユーザの興味・既知情報・直前の対話文脈などのコンテキストを推定し，プロンプトに反映する仕組みが課題となる．例えば SocialMind[3] は，音声・視線・ジェスチャ等のマルチモーダル情報を継続的に収集してユーザ状態を推定し，会話支援における有効性を示している．本研究でも同様に，ユーザコンテキストの導入により候補質問の網羅性向上が期待できる．ただし，候補提示に基づく選択操作は設計上，テキスト自由入力と比べて「ユーザが作り得る質問」を完全に網羅することは難しく，どの程度の網羅性を目標とし，どのように不足を補うか（再生成，候補編集など）の検討が必要である．また，本稿では音声入力を含めた AR グラス上の AI アシスタントへの文字入力を不要にし，視点画像に基づいて質問を生成・選択する手法自体を重視しし，遅延に関して実装では考慮しなかった．だが，実用化にはユーザの待機時

間の短縮のための入力画像の圧縮などの低遅延化策やプライバシー保護のためのオンデバイスでの推論，質問選択後の AI からの回答をユーザに伝えるインターフェイスの検討が必要になる．今後は，実環境にてユーザーが実際にしたい質問を選べるか，またその行為が外部から認識されやすいかをユーザ実験していく．

6. まとめ

本稿では，視点画像からマルチモーダル LLM が質問候補を生成し，ユーザがヘッドポインティングで選択する AR グラス向け AI アシスタントへの入力手法 LooQuestion を提案した．22 名の質問パターンに基づくプロンプト設計で 55.0%の一致率を得ており，専用学習なしでも一定の網羅性が確認できた．音声や大きなジェスチャに頼らないヘッドポインティング中心のインタフェースは日常的な AR 利用に向けて有望であり，ユーザのコンテキストを推論システムに埋め込み網羅性を上げるのが課題となる．XREAL ハードウェア上で実装可能であることを示したうえで，今後は実環境での利用者評価を通じて対象多様性へのカバレッジを検証する予定である．

参考文献

[1] Xu, F., Zhou, T., Nguyen, T., Bao, H., Lin, C. and Du, J.: Integrating augmented reality and LLM for enhanced cognitive support in critical audio communications, *International Journal of Human-Computer Studies*, Vol. 194, p. 103402 (online), DOI: https://doi.org/10.1016/j.ijhcs.2024.103402 (2025).

[2] Zhuang, K., Huang, Z., Song, Y., Li, R., Zhou, Y. and Yang, A. Y.: 2024 NASA SUITS Report: LLM-Driven Immersive Augmented Reality User Interface for Robotics and Space Exploration, (online), available from ⟨https://arxiv.org/abs/2507.01206⟩ (2025).

[3] Yang, B., Guo, Y., Xu, L., Yan, Z., Chen, H., Xing, G. and Jiang, X.: SocialMind: LLM-based Proactive AR Social Assistive System with Human-like Perception for In-situ Live Interactions, (online), available from ⟨https://arxiv.org/abs/2412.04036⟩ (2024).

[4] Shen, K., Wu, L., Tang, S., Xu, F., Long, B., Zhuang, Y. and Pei, J.: Ask Questions With Double Hints: Visual Question Generation With Answer-Awareness and Region-Reference, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 46, No. 12, p. 9648–9660 (online), DOI: 10.1109/tpami.2024.3425222 (2024).

[5] Mi, L., Montariol, S., Castillo-Navarro, J., Dai, X., Bosselut, A. and Tuia, D.: ConVQG: Contrastive Visual Question Generation with Multimodal Guidance, (online), available from ⟨https://arxiv.org/abs/2402.12846⟩ (2024).

[6] Xie, J., Zheng, J., Fang, W., Cai, Y. and Li, Q.: Explicitly diverse visual question generation, *Neural Networks*, Vol. 184, p. 107002 (online), DOI: https://doi.org/10.1016/j.neunet.2024.107002 (2025).

[7] Mulla, N. and Gharpure, P.: Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications, *Prog Artif Intell 12*, (online), available from ⟨https://doi.org/10.1007/s13748-023-00295-9⟩ (2023).

[8] Li, Y., Wang, L., Hu, B., Chen, X., Zhong, W., Lyu, C., Wang, W. and Zhang, M.: A Comprehensive Evaluation of GPT-4V on Knowledge-Intensive Visual Question Answering, (online), available from ⟨https://arxiv.org/abs/2311.07536⟩ (2024).

[9] Mohapatra, P., Aroudi, A., Kumar, A. and Khaleghimeybodi, M.: Non-verbal Hands-free Control for Smart Glasses using Teeth Clicks, (online), available from ⟨https://arxiv.org/abs/2408.11346⟩ (2024).

[10] Igarashi, Y., Futami, K. and Murao, K.: Silent Speech Eyewear Interface: Silent Speech Recognition Method Using Eyewear and an Ear-Mounted Microphone with Infrared Distance Sensors, *Sensors (Basel)*, Vol. 24, No. 22, p. 7368 (online), DOI: 10.3390/s24227368 (2024). Published Nov 19, 2024.

[11] Fang, F., Zhang, H., Zhan, L., Guo, S., Zhang, M., Lin, J., Qin, Y. and Fu, H.: Handwriting Velcro: Endowing AR Glasses with Personalized and Posture-adaptive Text Input Using Flexible Touch Sensor, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Vol. 6, No. 4 (online), DOI: 10.1145/3569461 (2023).

[12] Yuasa, R. and Nagao, K.: FlickPose: A Hand Tracking-Based Text Input System for Mobile Users Wearing Smart Glasses, *Applied Sciences*, Vol. 15, No. 15 (online), DOI: 10.3390/app15158122 (2025).

[13] Lee, J., Wang, J., Brown, E., Chu, L., Rodriguez, S. S. and Froehlich, J. E.: GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality, *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, (online), DOI: 10.1145/3613904.3642230 (2024).

[14] Blattgerste, J., Renner, P. and Pfeiffer, T.: Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views, *Proceedings of the Workshop on Communication by Gaze Interaction*, COGAIN '18, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3206343.3206349 (2018).

[15] Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X. and Vanderwende, L.: Generating Natural Questions About an Image, (online), available from ⟨https://arxiv.org/abs/1603.06059⟩ (2016).