

マルチラウンド明確化メカニズムによる Visual Question Answering における 曖昧質問応答の改善

Improving Ambiguous Question Answering via a Multi-Round Clarification Mechanism for Visual Question Answering

王 洋 (Wang Yang)*

法政大学 情報科学研究科 情報科学専攻 24T0005
yang.wang.6g@stu.hosei.ac.jp

Abstract

This study explores ambiguous question answering in multi-modal visual dialogue, where users interact with a system through multi-round questions grounded in images. In visually complex and semantically uncertain scenarios, users often ask ambiguous questions, which can lead to system misunderstanding. To address this problem, we propose a unified framework for ambiguity detection and clarification question generation. The framework first detects question ambiguity by combining answer distribution entropy and visual-text semantic similarity. To support targeted clarification, the system constructs a belief state that represents the visual context through object-level attribute information. Based on this belief state, a large language model generates context-aware clarification questions over multiple dialogue turns, enabling the system to progressively resolve ambiguity through interaction. Experiments on the SIMMC 2.0 dataset demonstrate that the proposed method effectively detects ambiguous questions. Additional evaluation on a self-collected supermarket shopping dataset shows that belief state-based clarification generation significantly improves the relevance and effectiveness of clarification questions. Overall, the proposed framework provides a complete multi-turn pipeline for ambiguity detection, belief state tracking, and natural clarification dialogue generation in multi-modal visual scenarios.

1 はじめに

近年、人工知能およびコンピュータビジョン技術の急速な発展に伴い、Visual Question Answering(VQA)[14] は、インテリジェントアシスタント、人間とコンピュータのインタラク

ション、サービスロボットなどの分野で広く注目されている。VQA システムは、画像中の視覚情報を理解し、自然言語を用いてユーザと対話することを目的としている。しかし、既存研究の多くは明確な質問への回答に重点を置いており、現実環境で頻繁に発生するあいまいな指示や歪んだ質問への対応については十分に検討されていない。特に、現実世界とのリアルタイムな相互作用を伴うシナリオにおいて、この問題はより顕著である。

従来の電子商取引システムでは、ユーザは構造化された商品カタログを通じて商品を閲覧し、クリックやドラッグ操作により選択および購入を行う。このようなインタラクションは完全にデジタル化された環境に基づいており、各商品は独立したインターフェース要素として提示され、操作は直感的かつ効率的である。一方、リモートショッピングロボットは、現実のスーパーマーケット環境を対象としており、ユーザはロボットのカメラを通じて商品を観察し、自然言語の指示で操作を行う。このインタラクションモードでは、ユーザは友人と一緒に買い物をするように、ロボットと対話しながらリアルタイムで目標商品を確認・選択でき、操作の自然性および正確性を向上させる。また、本システムは没入型の体験を提供し、商品配置や属性を認知することで環境への参加感を高める。操作過程では複雑なインターフェースに依存せず、ユーザの両手を解放することで、利便性と操作の流畅性を向上させる。

しかし、実際のシナリオにおいて、ユーザの自然言語による指示はしばしばあいまいである。例えば、「リンゴを買いたい」や「緑色のものは何か」といった指示は、複数の外観が類似した物体が存在する場合に容易に歪みを生じる。このため、複雑な視覚環境下のインタラクションシステムは、二つの主要な課題に直面する。第一に、目標物体認識の不確実性：シーンには複数の類似または繰り返しの物体、遮蔽、背景ノイズが含まれることがある。第二に、ユーザ意図理解の不確実性：同一画像に対して、主観性により複数の妥当な回答が存在する可能性がある [10]。このような課題により、単一ラウンドの質問応答の

* 指導教員：伊藤克亘 教授

みではタスク完了を保証できず、システムはマルチラウンド対話を通じて明確化質問を能動的に生成し、ユーザの真の意図を逐次的に特定する必要がある。

上述の問題に対応するため、本研究ではリモートスーパーマーケットシナリオにおいて、物体属性、あいまい性ラベル、マルチラウンドの明確化対話を含む自作データセットを構築し、複雑な現実環境下での視覚対話システムの性能を検証した。また、あいまい性検出および複数解答生成の実験には、SIMMC2.0 データセットおよび VQA データセットも併用した。自作データセット上の実験では、商品種別の多様性、陳列の乱雑さ、頻繁な遮蔽により、最先端の視覚モデルでも認識精度が低下することが確認された。これに対処するため、OCR モジュールを導入し、商品パッケージのテキスト情報と視覚特徴を融合させたマルチモーダル認識を実現した。自作データセット上で、この手法は物体認識精度を約 70% に向上させ、ピルーフステートの構築およびマルチラウンドの明確化質問生成に信頼性の高い基盤を提供した。

これらのデータおよび手法に基づき、本研究では多ラウンドの視覚対話とピルーフステート更新を可能とするシステムを構築した。これにより、複雑な現実環境下でのあいまいな指示の解消、インタラクションの自然性向上、没入型リモートショッピング体験の実現に技術的な支援を提供することが可能となる。

2 関連研究

本研究は、VQA およびビジュアルダイアログシステムの発展と密接に関連している。関連研究は大きく 3 つの側面に分類できる：(1) 視覚と言語の特徴抽出およびマルチモーダル融合、(2) あいまい性の認識、(3) 明確化質問 (clarification question) の生成である。

2.1 特徴抽出とマルチモーダル融合

視覚特徴抽出の分野では、Detic[15] は、未注釈カテゴリの物体を認識可能にし、大規模な画像レベル監督を利用することでバウンディングボックス予測の品質も向上させ、複雑なシーンにおける多様な物体の局所化精度を高めることができる。また、PaddleOCR[2] を用いて製品パッケージ上のテキスト情報を抽出し、視覚情報と融合することでマルチモーダル認識を実現した。

言語特徴抽出については、BERT[7] が双方向 Transformer エンコーダを採用し、大規模コーパスで事前学習された文脈依存の言語表現を獲得する。双方向性により、単語が前後の文脈に基づいて解釈され、意味の一貫性やあいまい性検出に有用な基盤を提供する。

視覚と言語の融合に関しては、ViLBERT[5] が二重ストリームのエンコーダ構造を採用し、共注意 Transformer によって視覚と言語特徴の細粒度な相互作用を実現する。しかし、このモデルの主目的は単一の回答を予測することであり、不確実性の検出を想定していない。それでも、VQA における高い性能

は、本研究においてユーザ質問と画像特徴の融合の参考となっている。

2.2 曖昧性の認識

質問のあいまい性定義は、既存研究では大きく 2 種類に分類される。1 つ目は、統語的・意味的あいまい性などの言語的あいまい性であり [8]、字義的解釈と語用論的解釈の不一致により生じることが多い [1]。2 つ目は、代名詞や省略表現の指示対象が不明確な指示的あいまい性である [9]。意味的あいまい性について、Terao[6] は予測された回答分布のエントロピーを計算することであいまい性を測定する手法を提案した。エントロピーが高いことは、複数の妥当な回答の間でモデルが不確実であることを示し、あいまい性の存在を意味する。指示的あいまい性について、Guo[4] らは、シーン内に視覚的に類似した物体が多数存在する場合、参照の曖昧さが生じやすいことを指摘している。さらに Bhattacharya[11] らは、画像自体が質問に答えるのに十分な情報を含まない場合、モデルが言語的手がかりのみに頼らざるを得ないケースがあることを観察した。また、あいまい性検出の一部の研究は、あいまい性の原因説明を目的としておらず、エンドツーエンドの教師あり学習によって直接ラベルを学習する。たとえば SIMMC 2.0[13] では、マルチモーダル対話に曖昧性判定タスクを導入し、人手によるラベルを用いて質問が曖昧であるか否かを学習させる。これに対し本研究は、回答分布のエントロピーと視覚-言語の類似度比較という 2 つのシグナルに基づく教師なしのあいまい性推論方法を提案し、言語モダリティと視覚モダリティの両方における不確実性を捉える。SIMMC のような教師あり手法と異なり、本手法は注釈付きデータを必要としない。

2.3 明確化質問生成

近年、大規模言語モデル (LLM) はオープンドメインの意味理解と自然言語生成において顕著な能力を示し、明確化質問を自動生成することが可能になった。適切なプロンプト設計により、LLM は多様で自然なフォローアップ質問を生成でき、ユーザ意図の理解を向上させることが先行研究で示されている。たとえば AGENTCQ システム [3] は、プロンプトによって LLM を誘導して明確化質問を生成し、対話システムにおけるクエリ解釈を改善する。

しかし、既存手法には依然としていくつかの制約がある。一方では、明示的な視覚的制約が存在しないため、生成される確認質問が特定の視覚的実体と正確に対応しない可能性がある。もう一方では、マルチラウンド対話においてユーザ意図の変化を表すピルーフステート [13] が十分に活用されておらず、意図の更新を動的に追跡することが困難である。その結果、生成される明確化質問の自然さや有効性は、複雑な視覚シーンや参照のあいまい性が存在する場合に制約されやすい。

これらの課題に対処するため、本研究では LLM に基づく明確化質問生成と、マルチラウンドのピルーフステート更新を統合した手法を提案する。

3 提案手法

本研究では、ユーザの質問と視覚シーンからあいまい性を自動的に識別し、明確化対話を生成するマルチランド対話システムを開発する。提案システムは図1に示すように、2つの中核モジュールからなるクローズドループ構造を持つ。第1のあいまい性検出モジュールは、回答分布のエントロピーと画像・言語間の意味類似度を統合して質問の曖昧さを判定する。曖昧と判断された場合、第2の明確化対話生成モジュールが起動し、LLMを用いて自然な明確化質問を生成する。システムは情報利得と意味的関連性に基づき最適な物体属性を選択し、ユーザのフィードバックに応じてピリーフステートを動的に更新する。この更新されたピリーフステートは次の明確化対話へ再入力され、反復的なインタラクションループを形成する。

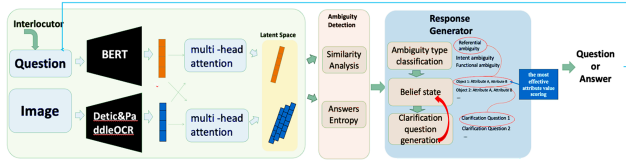


図1: システム全体のアーキテクチャ

3.1 視覚的意味曖昧性検出モジュール

本研究では視覚-言語類似度と回答分布エントロピーに基づく非教師あり曖昧性検出手法を提案する。

3.1.1 視覚-言語融合による候補回答分布の生成

まず、ViLBERTを基盤とする視覚-言語融合ネットワークを用い、DeticおよびOCRによって抽出された視覚特徴と、BERTによって得られた言語特徴を統合し、候補回答に対する確率分布を生成する。

与えられた画像 I と質問 Q に対して、視覚特徴および言語特徴をそれぞれ次のように抽出する：

$$\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}, \quad \mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_n\}. \quad (1)$$

これらの特徴をマルチヘッド注意機構 (Multi-Head Attention) により融合し、統合特徴表現 \mathbf{F}_i を得る。

$$\mathbf{F}_i = \text{MultiHeadAttn}(\mathbf{V}, \mathbf{Q}). \quad (2)$$

その後、分類ヘッドを用いて、各候補回答 a_k に対する確率 A_k を出力する：

$$A_k = \sigma(Z_k) = \frac{1}{1 + \exp(-Z_k)} \quad (3)$$

ここで、 σ は Sigmoid 関数であり、 Z_k は候補回答 a_k に対応する logit 値を示す。次に、ユーザー質問の指示的明確性を評価するため、質問表現 \mathbf{q}_i と画像中の各候補物体表現 \mathbf{v}_i との意味的類似度を計算する：

$$\text{sim}_i = \cos(\mathbf{q}_i, \mathbf{v}_i). \quad (4)$$

すべての物体における最大類似度 $\max(\text{sim}_i)$ が閾値 τ_{sim} より低い場合、質問と視覚実体との対応が弱く、潜在的な曖昧性が存在すると判断する。

一方、モデル出力の回答分布はシステムの不確実性を反映している。複数の候補回答に対する信頼度が近い場合、質問が意味的に曖昧である可能性が高い。そのため、回答分布のエントロピーを導入して不確実性を定量化する：

$$H(\mathbf{a}) = - \sum_{i=1}^n a_i \log a_i. \quad (5)$$

ここで、 a_i は第 i 番目の候補回答に対応する正規化確率を示す。エントロピー $H(\mathbf{a})$ が閾値 τ_{entropy} を超える場合、モデルは回答予測に高い不確実性を示し、曖昧な質問であると判断する。

3.2 明確化対話生成モジュール

従来のテンプレートに基づく質問生成手法とは異なり、本研究ではまず、入力質問とシーン内の候補物体とのマッチング特徴に基づいて、質問に含まれる曖昧性の種類 (指示的曖昧性、属性的曖昧性、語義的多義性など) を判定する。その後、システムは予測された曖昧性タイプを制御信号として利用し、物体属性を意味的制約として組み込み、大規模言語モデル (LLM) の生成能力を用いて、曖昧性検出とピリーフステート更新を統合し、多段階の対話を通じて不確実性を段階的に解消していく。

3.2.1 曖昧性タイプ判定

語義的曖昧性は、多義的な指示、欠落している属性情報、あるいはユーザ意図の不明瞭さなどから生じ得る。本研究では曖昧性を、指示的曖昧性、属性的曖昧性、意図的曖昧性、用途的曖昧性の四種類に分類し、マルチモーダル分類器が入力サンプルを自動的にこれらのタイプに分類し、その結果を明確化質問生成の条件として用いる。

3.2.2 属性抽出とピリーフステートモデリング

マルチモーダル環境において、システムは物体を認識するだけでなく、その意味属性およびそれらがユーザ意図とどのように関係するかを理解する必要がある。本研究では、視覚情報、LLMに基づく推論、および対話コンテキストを統合し、各物体およびその属性値に初期確信度を付与することで、初期ピリーフステート B_0 を構築する。本研究では属性を三つの階層レベルに分類する。視覚レベル属性 (色、形状、位置、大きさ、価格など) は視覚特徴から直接抽出される。用途レベル属性 (「食用可能」「加熱可能」「着用可能」「清掃用」など) は、物体の機能や用途を記述するものであり、画像から直接取得できないため、LLMによって自動生成される。意図レベル属性 (「購入目的」「贈答用」「保存用途」「試用目的」など) は、ユーザの意図に関わる属性である。マルチラウンド t において、ピリーフステート B_t は次のように形式化される：

$$B_t = \{o_i, P(o_i), \{a_{i1}, a_{i2}, \dots, a_{im}\}\}. \quad (6)$$

ここで、 o_i は i 番目の物体を表し、 $P(o_i)$ は物体 o_i が

元の質問に関連すると判断される確信度を示す。また、 $\{a_{i1}, a_{i2}, \dots, a_{im}\}$ は物体 o_i に関連付けられた属性集合である。ピリーフステートは、現在の候補物体に関する属性および文脈理解を保持する動的な知識構造を表す。例えば、ユーザが「これは節電に役立ちますか?」と質問した場合、システムは視覚認識を行い、シーン内に二つの類似したランプ（スマートランプと装飾用ランプ）が存在することを特定する。初期ピリーフステートは、用途属性として「省エネ機能」: 0.6、「装飾機能」: 0.4 を付与する。ユーザが「このスマートランプのことです」と回答した場合、「省エネ機能」の確信度が上昇し、「装飾機能」の確信度は低下する。

3.2.3 候補属性値のスコアリング

曖昧性解消のための明確化質問生成を効果的に行うには、システムは候補物体の中でどの属性値が最も識別に有効であるかを判断する必要がある。複数の候補物体が存在する場合、すべての属性値が同じように有用であるわけではない。ある属性値は多くの候補に共通して現れる（例：色属性における「赤」）一方で、別の属性値は希少で識別性が高い（例：チーズ選択における「マイルドな味」）。これを定量化するために、本研究では属性値 v_j の情報量を評価する指標として、まず候補物体間での頻度を計算する：属性値 v_j の頻度は、その値を含む候補物体の割合として定義される：

$$\text{frequency}(v_j) = \frac{|\{o \in C_t \mid v_j \in o\}|}{|C_t|}. \quad (7)$$

ここで、 C_t はマルチラウンド t における候補物体集合を示す。頻度が高いほど、多くの候補物体が同一の属性値を共有しており、その属性値は物体間の識別には不向きであることを意味する。

この直観を形式化するために、属性値の情報利得を次のように定義することで、曖昧性解消対話を可能にする：

$$\text{gain}(v_j) = \frac{1}{\text{frequency}(v_j)}. \quad (8)$$

しかし、識別力のみでは十分ではない。システムはさらに、属性値の意味的関連性を評価する。これは、属性値がユーザ質問、ユーザフィードバック、および現在の対話文脈とどの程度整合しているかを測る指標であり、次のように定義される：

$$\text{relevance}(v_j) = f_{\text{semantic}}(v_j, Q, A_{\text{user}}, C_t). \quad (9)$$

ここで、 Q は初期ユーザ質問、 A_{user} はユーザの過去の応答、 C_t はマルチラウンド t までの対話文脈を表す。本研究では、 f_{semantic} を埋め込みベースの類似度として実装し、属性値テキストと文脈テキスト（質問およびフィードバック）を同一の埋め込み空間にエンコードした上で、コサイン類似度を関連性スコアとして用いる。

最後に、各属性値の総合スコアは次のように計算される：

$$\text{score}(v_j) = \text{gain}(v_j) \times \text{relevance}(v_j). \quad (10)$$

この指標により、情報量 (gain) と文脈適合性 (relevance) の両方が調和され、各対話ラウンドにおいて最も有効な明確化焦点 v_j を選択することが可能となる。

3.2.4 明確化質問生成

曖昧性タイプの識別と属性抽出が完了すると、システムは明確化質問生成段階へと進む。従来のテンプレート依存型の手法とは異なり、本研究では意味制御に基づく条件付き生成方式を採用し、大規模言語モデル (LLM) の文脈理解および自然言語生成能力を利用して、各種曖昧性タイプに応じた自律的な応答生成を実現する。具体的には、明確化質問生成モジュールへの入力は、ユーザの元質問 q_{user} 、曖昧性タイプラベル t 、現在のピリーフステート B 、選択された属性値 v_j 、および対話文脈 C から構成される。これらの制約情報は LLM の入力プロンプトに埋め込まれ、次のように形式化される：

$$Q_{\text{clar}} = \text{LLM}(q_{\text{user}}, t, B, v_j, C). \quad (11)$$

条件付き明確化質問生成において、LLM は以下の二つの要件を満たす質問を生成する：(1) 意味的自然性：人間が自然に発話する質問のように流暢であること；(2) 曖昧性解消の適合性：最も識別力が高く、かつ文脈と整合する属性値 v_j に基づいて、ユーザ意図を効果的に明確化すること。

例として、以下の入力を与えた場合を考える。

User's question: "Does this help me save electricity?"
 Scene: Two objects detected — Smart Lamp (object A) and Decorative Lamp (object B)
 Ambiguity type: usage
 Task: Generate a natural question to clarify the intended function.

このとき、「省エネ機能」が最も有効な属性値 v_j として選択された場合、LLM は次のような明確化質問を生成することがある（例）：

"Clarification question": "Are you referring to the smart lamp or the decorative lamp next to it?",
 "Focus attribute": "function"}

ユーザの応答後、システムはピリーフステート内の該当属性スロットの確信度分布を動的に更新し、次ラウンドにおけるより精密な意味参照解決とユーザ意図推定を可能にする。このような意味条件付き明確化質問生成によって、モデルはさまざまな曖昧性シナリオにおいて自律的にフォローアップ質問を生成し、意味理解と対話生成を統合した連続的な知的振る舞いを実現する。

4 応用例

本研究で提案した手法に基づき、適用可能なシナリオを設計した。身体の不自由な人が自宅からスーパーでの買い物体験をできるようにする：典型的なシナリオでは、人間がロボットのカメラを通して操作すると、スーパーの一部しか見ることがで

きない。他のエリアを確認するためには、「左に回して」「右に回して」といった指示を出したり、「左には何がありますか?」「右には何がありますか?」と尋ねたりする必要がある。これ煩雑で効率的ではなく、プロセスを合理化する知能が欠けている。

本研究では、ロボットが買い物中の友達のように振る舞い、商品の話題を共有して体験をより楽しいものにする(図2)。視覚対話学習を通じて、ロボットは自律的に物体を認識し、人間の意図を理解して、商品の推薦などの行動を自主的に行うことができる。例えば、人間がリンゴを買いたい場合、ロボットは自動的に果物売り場に移動し、赤や緑のリンゴ、異なる価格やパッケージ、数量などを見つける。その後、単に指示に従うのではなく、「どの種類のリンゴを購入したいですか?」と質問することができる。視覚的理解に基づいて明確化質問を行うことで、ロボットはより意味のある会話を促進し、絶えず指示を出す必要を減らしながら、買い物体験を向上させる。



図2: 提案手法の視覚対話システムによる遠隔スーパー買い物の利用シナリオ

5 実験

本章では、本研究で提案したモデルの性能を評価するため、曖昧質問の検出および明確化質問の生成に関する実験を行った。

5.1 曖昧質問検出実験

5.1.1 データセット

本研究では、マルチモーダル対話における質問の曖昧性検出のために SIMMC2.0 データセットを用いてモデルを学習した。SIMMC2.0 は実世界の買い物および家具選択シナリオに基づく対話データであり、約 11.2k 件の対話 (117k 発話) を含み、その約 3 分の 1 が曖昧な質問として注釈されている。また、複数回答生成の性能評価については、Visual Question Answering (VQA) タスク設定に基づいて検証を行った。

5.1.2 曖昧質問検出および回答生成

まず、テキスト特徴と画像特徴の類似度を比較し、類似度が低い場合には質問を曖昧であると判断する。次に、画像特徴とテキスト特徴を融合し、線形画像を通じて回答空間へ投影することで、一つまたは複数の候補回答を生成する。各候

補回答に対して Sigmoid 関数を適用し、得られた回答分布のエントロピーを計算することで、モデルの不確実性を定量化する。実験では、曖昧性検出の閾値を以下のように設定した。類似度に基づく閾値は 0.2、回答分布のエントロピーに基づく閾値は 1.7 とした。最大類似度が 0.2 未満であり、かつエントロピーが 1.7 を超える場合には、類似度とエントロピーの両方を用いた統合的判断戦略を適用する。これらの閾値設定により、多段階の曖昧性検出が可能となる。学習は 2 枚の GPU (Titan X および Titan V を使い、バッチサイズ 128, エポック数 20 で実施した。最適化手法には Adam を使い、初期学習率は 1×10^{-4} に設定した。

5.1.3 実験結果および分析

タスク 1: 複数回答候補予測。 曖昧な視覚シーンにおいて、質問-画像ペアに対する複数の妥当な回答を予測するモデル能力を評価する。表 1 に示すように、本モデルは SIMMC2.0 において MM-SDT ベースラインを上回り、精度 0.43 から 0.54 へと向上させた。また、本モデルは VQA v2.0 において 0.78 の精度を達成しており、異なる視覚質問応答ベンチマークに対する高い汎化性能を示している。さらに、図 3 として、モデルが生成した複数回答の具体例を示す。

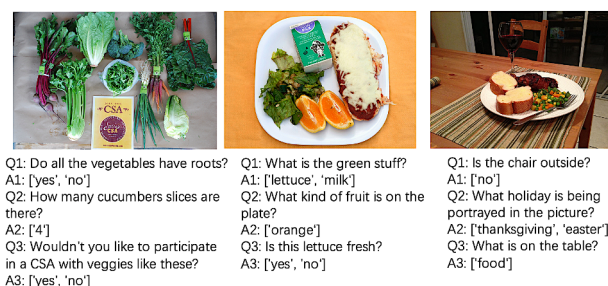


図3: 複数解答候補予測

表 1: 複数回答における予測精度

Model	Dataset	Accuracy
MM-SDT (SIMMC)	SIMMC 2.0	0.43
Our Model	SIMMC 2.0	0.54↑
Our Model	VQA v2.0	0.78↑

タスク 2: 曖昧性分類。 本タスクは、ユーザーの質問が曖昧であるか否かを識別することを目的とする。曖昧性検出のために、回答分布のエントロピー、視覚-言語間の意味的類似度、および両者を統合した特徴を用いる。表 2 に示すように、エントロピーに基づく特徴は 0.762 の分類精度を達成し、GPT2 ベースラインを上回る性能を示した。一方、類似度に基づく特徴は精度ではやや劣るものの、F1 スコアにおいては最も高い 0.550 を達成しており、曖昧な事例を捉える能力に優れていることが示唆される。さらに、エントロピーと類似度の両特徴を統合することで、精度および F1 の両指標においてより balan

スの取れた性能が得られ、モデルの頑健性が向上する。実運用においては、明確化対話を開始する際の精度と再現率のバランスを考慮し、本システムでは統合モデルを採用している。図4として、問題-画像ペアに対する具体的な曖昧性判断例も示す。

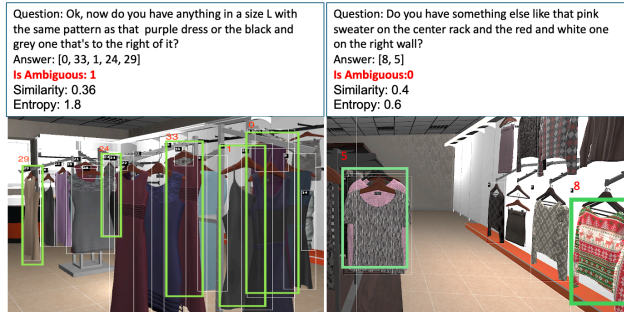


図 4: 曖昧な質問の検出

表 2: SIMMC2.0 での曖昧性解消の分類精度

Model	Accuracy	F1
GPT2 (SIMMC)	0.43	-
Entropy-based (Our)	0.762↑	0.459
Similarity-based (Our)	0.450	0.550
Entropy + Similarity Fusion (Our)	0.550	0.490

5.2 明確化質問生成に関する実験

5.2.1 データセット:

入力データとして、実際のスーパーマーケットでの買い物場面を模擬した独自データセットを用いた。本データセットは 350 枚の購買シーン画像、100 名の被験者から収集した 1,050 件の質問および 5,250 件の回答から構成されている。画像中の物体属性（色、形状、機能など）は Detic および PaddleOCR を用いて抽出し、ベリフステートとしてモデルに入力した。提案システムは物体属性を自動抽出する能力を有しているが、本実験では入力の一貫性を保証し、ベリフステート条件付けが明確化質問生成に与える影響を明確に分析するため、すべての実験において手動で指定したシーン属性を使用した。

5.2.2 モデルおよび入力処理

明確化質問生成のための大規模言語モデルとして、Mistral-7B-Instruct-v0.3 をバックボーンモデルとして採用した。モデルへの入力には、ユーザの質問 (q_{user})、ベリフステート (B)、およびシーン情報 (scene) を含めた。本研究では、プロンプト駆動型生成手法を採用し、ベリフステート表現を段階的に組み込むことで、異なる条件下における明確化質問生成能力を比較評価した。

5.2.3 定性的評価およびケーススタディ

実験設定の妥当性を検証し、実運用におけるモデルの振る舞いを示すために、異なる種類のチーズを含む実際のスーパーマーケット場面を用いた明確化質問生成の具体例を提示する。

本対話シーンは図 4(b) に示される画像に対応しており、視覚的に 6 種類のチーズ製品が含まれている。ユーザの質問は「ピザ用のチーズを買いたいのですが、どれを選べばよいですか?」である。この質問は、複数の候補チーズがユーザの意図を満たし得るため、使用目的に基づく曖昧性を含んでいる。



図 5: 実験シーンの例

このような状況において、モデルはまず各チーズ製品に対するベリフステートを生成し（表 3 参照）、その結果を基に曖昧性分析および明確化質問生成を行う。

表 3: ベリフステート (Belief State) 生成のプロセスと形式

Prompt: You are an assistant that generates a complete belief state for all objects in a scene.
User question: I want to buy cheese for pizza. Which one should I choose?
Scene items: [Red Cheddar 100g 499 yen; Cream Cheese 200g 389 yen; Kiri Creamy Portion 6p 359 yen; Mozzarella 349 yen; Tokachi Camembert 399 yen; Munster Cheese 90g 498 yen]
Ambiguity type: usage
Output example:
[{"type": "Red Cheddar", "color": "Red", "position": "1", "quantity": "100g", "price": "499 yen", "intended use": ["pizza", "sandwich", "snack"], "use method": ["grating", "slicing"], "serving context": ["main dish", "side dish"], "additional attributes": ["melts easily", "sharp flavor", "firm texture", "common choice for pizza"], "notes": "Popular for pizza due to its strong flavor and melting properties"}]

5.2.4 属性選択とピリーフステートの更新

各対話ラウンドで候補物体の属性分布を推定し、情報利得と意味的関連度を統合して各属性値のスコアを算出する。その中から、候補物体を最も識別しやすい属性を明確化質問の焦点属性 V_j として選択する (Tables 4 および 5 参照)。第 1 ラウンドでは、「mild flavor (マイルドな風味)」が選択され、Mozzarella (マイルド) と Red Cheddar (濃厚) を効果的に区別することに成功した。続く第 2 ラウンドでは、ユーザーからのフィードバックを反映した結果、「strong flavor (濃厚な風味)」が選択され、対象はさらに Red Cheddar に収束した。最終的に、目的物体の信頼度は所定の閾値 (例: 0.95) に到達し、他の候補は対話の進行とともに段階的に低下する。図 6 は、6 種類のチーズ候補に対するピリーフステート信頼度の動的変化を示している。

表 4: 第 1 ラウンドの候補属性値と統合スコア

Attr.	Attr. Value	Freq.	Gain	Relevance	Score
Serving context	Main dish	3	0.3333	0.49	0.1632
Addi. attr.	Melts easily	1	1.0000	0.37	0.3741
Addi. attr.	Sharp flavor	1	1.0000	0.36	0.3597
Addi. attr.	Common choice for pizza	2	0.5000	0.71	0.3571
Addi. attr.	Mild flavor	1	1.0000	0.40	0.4000

表 5: 第 2 ラウンドの候補属性値と統合スコア。

Attr.	Attr. Value	Freq.	Gain	Relevance	Score
Addi. attr.	Strong flavor	2	0.5000	0.38	0.1883

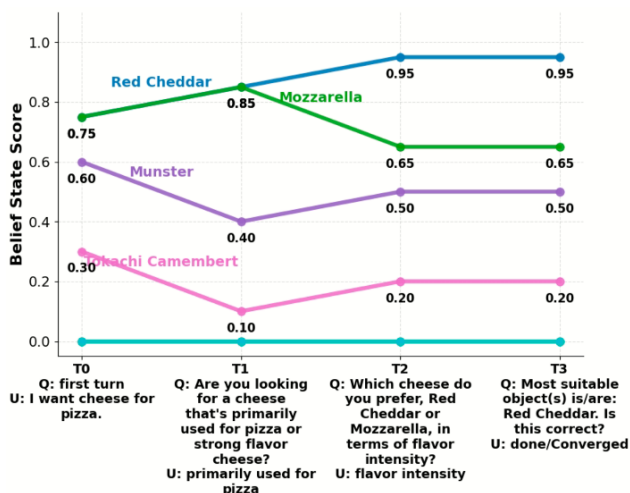


図 6: 対話ラウンド間におけるピリーフステートの変化

5.2.5 マルチラウンド明確化質問生成の結果

表 6 に示すように、ユーザーの初期質問は「ピザ用のチーズを購入したい」というものであり、複数のチーズが該当し得るため、本質的に曖昧性を含んでいる。提案手法では、各候補物体に対するピリーフステートを対話の進行に応じて動的に更新し、マルチラウンドの明確化対話を通じてユーザーの真の意図に段階的に焦点を当てていく。その結果、最終的に Red Cheddar が選択された。この選択は、ピザ調理に適しているという機能的要件を満たすだけでなく、ユーザーが好む「濃厚な風味」という嗜好とも一致しており、本システムが曖昧な要求を効果的に解消し、ユーザーニーズを的確に理解できていることを示している。単一選択に基づく単純な手法と比較して、本手法は対話的推論を通じてより柔軟で個別化された意思決定を可能にし、その実用性および研究的意義を十分に反映している。

表 6: 複数種類の曖昧性を対象とした明確化質問生成の例

Turn	Utterance
Prompt	You are a conversational AI system whose role is to generate natural and effective clarification questions for ambiguous user queries.
T0 (User)	I want to buy cheese for pizza. Which one should I choose?
T0 (System)	Are you looking for a cheese that's primarily used for pizza, like Red Cheddar or Mozzarella, or are you interested in a different type of cheese with a strong flavor, such as Munster Cheese?
T1 (User)	Primarily used for pizza.
T1 (System)	Which cheese do you prefer, Red Cheddar or Mozzarella, in terms of flavor intensity?
T2 (User)	Flavor intensity.
T2 (System)	Based on your question and feedback, the most suitable object(s) is/are: Red Cheddar, Mozzarella. Is this correct, or should I tell you more about the options?
T3 (User)	Done.

上記の結果から、手動でアノテーションされたシーン情報は、自然な明確化質問を生成する上で非常に有効であることが確認できる。比較のために、本研究ではスーパーマーケット画像から視覚レベルの属性を自動生成する実験も行った。具体的には、Detic を用いて物体検出を行い、VinVL[12] により属性を抽出し、さらに OCR を用いてテキスト情報を補完した。この自動処理によって生成されたシーン情報の例を Table 7 に示す。しかし、自動認識手法は、文字の重なり、あるいは遮蔽が生じている場合に誤認識が発生しやすく、ノイズを含む可能

性がある。その結果、明確化質問生成の挙動分析に影響を与える恐れがある。

表 7: シーン内アイテム認識におけるの出力結果 (Detic + VinVL + OCR)

Detic & OCR & VinVL outputs: Grand Alheirn 329 yen 356 R, item CS Tokexu-CoC 269291 YiO9, Tokeru Cheese Co6 499 539 U, company product 0 IRSV Soli Cheese coc vO9 269 291, Cream Portion 4 359 yen 97.8g 308, Tokere Cheese WMO 22”, Tsuru Ko Takanto 499 discount 1000 7E0P 530, Smoked Cheese CoC Y309 389 421, Camembert Cheese N-ku, Tokere Cheese Kome 100 229 50g 240, World Chitaars 097 1120279 302, Parmesan Cheese 29 N 499 yen 80g na539 JUPN., 12E4 CoC 6 Tchir Tai Fu 69 183 FAS, Rera Cheese 349 *377, Kiri 0 Cream Portion 4 359 yen 97.8g, Bisco Sauce CoC 2000 249.26 small Toru...
--

6 結論と今後の課題

本研究では、視覚と言語を統合したマルチモーダル環境における曖昧質問検知および明確化質問生成の課題に取り組んだ。

曖昧性検知に関しては、テキスト-画像間の意味的類似度と回答分布のエントロピーに基づく軽量なモデルを提案した。SIMMC2.0 データセットにおいて、本手法は複数回答予測で 0.54、曖昧性分類で 0.762 の精度を達成し、既存のベースライン手法を上回る性能を示した。

明確化質問生成に関しては、ピリーフステートに基づくモジュールを導入し、物体の多様で豊かな属性を明示的にモデル化することで、解釈可能な属性駆動型マルチランド対話を実現した。

今後の課題としては、自動画像認識によって生成されるシーン物体情報の品質向上に加え、より信頼性の高い自動認識条件下におけるモデルの適応性および拡張性の評価が挙げられる。また、生成され明確化質問の有効性を定量的に評価するための自動評価指標の開発にも取り組む予定である。

参考文献

- [1] W. Alex, S. Amanpreet, M. Julian, H. Felix, L. Omer, and B. Samuel. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop*, pages 353–355, 2018.
- [2] C. Can, S. Tao, L. Min, G. Jie, and M. Yao. Paddleocr 3.0 technical report. *arXiv:2507.05595*, 2025.
- [3] S. Cornelius, Y. Yifan, A. Mohammad, and de R. Maarten. Agent-cq: Automatic generation and evaluation of clarifying questions for conversational search with llms. *arXiv:2410.19692*, 2024.
- [4] G. Dandan et al. Gravl-bert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 285–297, 2022.
- [5] L. Jiasen et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [6] T. Kazuma et al. Rephrasing visual questions by specifying the entropy of the answer distribution. *IEICE Transactions on Information and Systems*, pages 2362–2370, 2020.
- [7] D. Jacob, C. Ming-Wei, L. Kenton, and T. Kristina. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- [8] C. Kenneth and P. Ramesh. Coping with syntactic ambiguity or how to put the block in the box on the table. *American Journal of Computational Linguistics*, 8(3–4):139–149, 1982.
- [9] P. Massimo and A. Ron. The reliability of anaphoric annotation reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, 2005.
- [10] B. Nilavra, L. Qing, and G. Danna. Why does a visual question have different answers? In *ICCV*, pages 4271–4280, 2019.
- [11] B. Nilavra, L. Qing, and G. Danna. Why does a visual question have different answers? In *ICCV*, pages 4271–4280, 2019.
- [12] Z. Pengchuan, L. Xiujun, H. Xiaowei, Y. Jianwei, and G. Jianfeng. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588, 2021.
- [13] K. Satwik, M. Seungwhan, G. Alborz, and D. Babak. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv:2104.08667*, 2021.
- [14] A. Stanislaw, A. Aishwarya, L. Jiasen, M. Margaret, B. Dhruv, Z. C. Lawrence, and P. Devi. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.
- [15] Z. Xingyi, G. Rohit, J. Armand, and M. Ishan. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, pages 350–368, 2022.