

# エージェントの発話とジェスチャによる 調理動画支援システムの構築

肥田 京佳<sup>1</sup> 徳久 良子<sup>1,2</sup>

**概要：**本稿では、調理動画における視聴者理解を支援するエージェント型インターフェースにおいて、エージェントの説明箇所を自動切り出しする手法を提案する。本手法は、調理場面におけるさまざまな食材や調理器具の位置・状態を言語情報と対応づけて解析し、レシピ文中の説明発話（例：「タマネギを微塵切りにします。」）と対応する画像内の領域（食材や調理器具など）を、GPT-4.1 を用いて抽出した上で、エージェント対話向けに補正する手法である。一般の被験者 120 名による評価実験の結果、提案手法の画像切り出しは、人間による画像切り出しと同等の精度であることが分かった。今後はエージェントの動作生成や、調理の学習に対する有効性の検証に取り組む予定である。

## A Cooking Video System Supported by an Agent with Utterances and Gestures

### 1. はじめに

インターネットにおける料理コンテンツは、テキストベースのレシピから動画形式へと移行しつつある。調理動画は、調理手順や動作を直感的に理解できる利点を持ち、学習者が自分のペースで繰り返し視聴できる点から、調理スキルの学習に有効であることが報告されている [1]。また近年では、こうした調理動画を対象として、映像・音声・テキスト情報を組み合わせて調理工程を解析するマルチモーダル手法 [2] や、調理動画をもとに調理手順を自動的に解析・生成する手法 [3], [4], [5] も提案されている。

一方で、調理動画には視聴者の理解を妨げるいくつかの課題も存在する。まず、動画には映像・音声・字幕といったさまざまな情報が同時に提示されるため、調理経験の少ない初心者にとっては「どこに注目すべきか」が分かりにくく、注意が分散しやすい。さらに、一般的な調理動画では視覚的な注意喚起が十分でなく、包丁の扱いや火の使用、油はねなどの危険な場面でも映像が淡々と進行するため、危険性が十分に伝わらない場合がある。その結果、視聴者が調理手順を誤って解釈したり、危険箇所を見落とす恐れがあり、調理に対する理解促進および安全性の観点から改



図 1: 提案システムの使用例（エージェントが指差しジェスチャを交えて補足説明している場面）

善が求められる。

そこで本研究では、料理番組のように補足や助言を行う映像支援システムの実現を目指す。図 1 に提案システムの動作例を示す。本研究の最終的な目的は、調理動画に対して動的な視覚的・音声的サポートを提供し、視聴者が安全かつ効果的に料理を学べる環境を構築することである。具体的には、エージェントが「ここを見てください」と視線を誘導したり、「しっかり炒めるとはどのような状態か」を補足説明したり、「包丁で手を切らないように注意してください」のように危険箇所を注意喚起する。このような支援を通じて、単にレシピを映像で伝達する調理動画を、調理手順や安全性の理解を深める学習支援コンテンツへと発展

<sup>1</sup> 愛知工業大学

<sup>2</sup> 理化学研究所

させることを目指す。本発表では、上述の調理支援動画システム構築に向けて、画像からユーザに注目してほしい箇所を切り出す手法について報告する。

## 2. 関連研究

### 2.1 COM-Kitchens データセット

COM-Kitchens は一般家庭のキッチンで撮影された大規模な調理動画データセットである [6]。すべての動画は、一般家庭の調理台に三脚を置いて設置したスマートフォンで真上から撮影されており、流し台やコンロなどキッチンの全体が映る構図となっている。動画には、切る、炒める、混ぜるといった複数の調理行為が含まれ、動作や食材の状態などがアノテーションされている。

本研究では、NII IDR で公開されている COM-Kitchens データセットを使用した<sup>\*1</sup>。公開版は、76 戸の一般家庭のキッチンで撮影された全 177 本の調理動画から構成され、157 種類のレシピを含む。これらの動画から静止画を抽出し、各調理ステップに対応する画像と説明発話のペアを作成した。説明発話は、COM-Kitchens に付与されていた調理動作のテキストを「～します」という形式に統一した上で、複数動作を含む文については動作ごとに文を分割して構築した。例えば、「タマネギを微塵切りし、ボウルに入れる。」という文は、「タマネギを微塵切りにします。」「タマネギをボウルに入れます。」のように整形した。また、動画内で実行されていても元のテキストに記載されていない動作（例：味噌を適量測ります。）については、動画を確認した上で人手で説明発話を追加した。

### 2.2 エージェントによるマルチモーダル対話研究

エージェントを用いたマルチモーダル対話の研究は、音声・視覚・身体動作を統合し、人間らしい対話の実現を目指して発展してきた。これらの研究では、発話内容に応じて表情・視線・ジェスチャなどの非言語的情報を生成し、ユーザの理解促進や親和性の向上を図っている。

アンドロイド型エージェントを用いた代表的な研究として、ERICA による傾聴対話システムが挙げられる [7]。この研究では、相槌や繰り返し、深掘り質問、評価応答などのさまざまな傾聴発話を生成する仕組みを構築し、人間との比較実験を通じて、エージェントによる自然な傾聴動作の有効性を検証している。実験の結果、自律的な対話制御によっても、手動操作と同程度に「話しやすい」「真面目に話を聞いていた」といった基本的な傾聴スキルは人間と遜色ないことが示されている。

また、MMDAgent-EX は、音声認識・音声合成・3D モデル表示・モーション制御を統合的に扱える対話型エージェント構築プラットフォームである [8]。本プラットフォーム

ムでは、FST (Finite State Transducer) ベースのスク립トを用いることで、発話と連動したモーション（例：手を振る、うなずく）を 3D モデルに割り当てただけでなく、複数モーションの組み合わせや状態遷移に応じた制御も可能となる。さらに、Python などのスク립ト言語と接続することで、外部の音声認識・音声合成エンジンや ChatGPT などのモジュールを組み込み、より複雑で柔軟な動作・対話を実現できる。ユーザが任意のモーションデータや音声、FST スクリプト等を作成することで、システム全体の拡張性を高めることができる。

本研究では、これら先行研究で示された「発話と非言語情報の統合による理解支援」の知見を踏まえ、MMDAgent-EX を用いて調理動画の発話内容と画像情報に基づくジェスチャ生成を行う。具体的には、「(野菜を) 切る」などの調理動作や動画の指差しなどを 3D キャラクタに動作させ、調理動画の視覚的補助を実現する。

### 2.3 CLIP

Contrastive Language-Image Pre-training (CLIP) [9] は、大規模な画像と言語のペアを用いて事前学習された視覚と言語の統合モデルである。画像とテキストの対応関係を学習し、その類似度を高精度に推定できる特徴を持つ。CLIP の特徴は、特定のタスクに特化した再学習を行わなくても、高い汎化性能を発揮する点にある。これにより、未知の画像や説明文に対しても既存の知識をもとに意味的対応づけを行うことができる。このような性質は、多様な食材や新しい調理手順が頻繁に登場する調理領域において特に有用である。本研究では、未知のレシピや食材に対しても柔軟に対応可能な視覚言語モデルとして、CLIP を用いた画像・テキストの意味的対応づけを行う。

## 3. 調理動画の支援システムの概要

本研究で提案する調理動画の支援システムの処理の流れを図 2 に示す。まず、入力として調理シーンの画像と説明発話（例：「タマネギを微塵切りにします。」）を受け取る。次に、図 2 の中央の枠内に示す通り、以下の 3 つのモジュールで処理を行う。

- (1) 画像切り出しモジュール：説明発話の内容に基づき、OpenAI の GPT-4.1 モデル [10] を用いて画像中の主要オブジェクト領域を検出し、切り出す（図 2 の例では、まな板の上のタマネギを検出し切り出している）。
- (2) ジェスチャ選択モジュール：説明発話と画像の内容に基づき、発話意図を補う補足説明を生成し、内容に適したジェスチャを選択する。例えば、「玉ねぎを切る」場面に対して「猫の手で指を守りましょう」という補足説明が得られた場合には、手を丸める動作を選択する。各ジェスチャには動作内容を示す説明文を定義しておき、CLIP により補足説明とのテキスト類似度と

<sup>\*1</sup> [https://www.nii.ac.jp/dsc/idr/rdata/COM\\_Kitchens/](https://www.nii.ac.jp/dsc/idr/rdata/COM_Kitchens/)

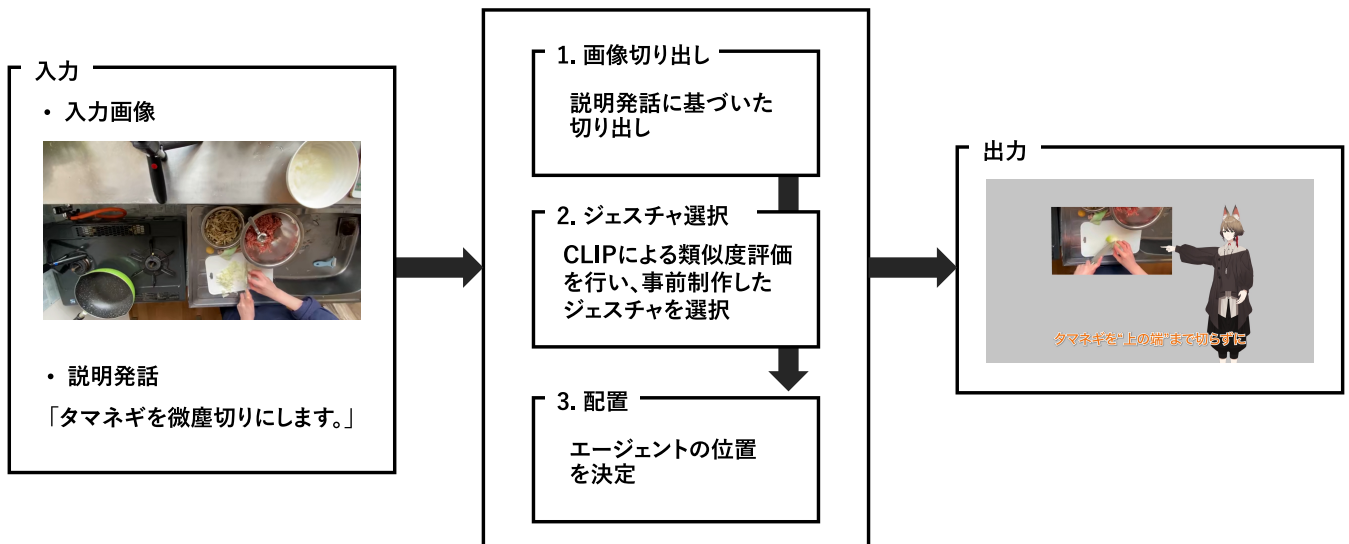


図 2: 本システムの流れ

切り出し画像との画像類似度を算出し、それらを重み付きで統合して最適なジェスチャを決定する。

(3) 配置モジュール：切り出した画像とエージェントの位置関係を考慮し、不必要な被写体との重なりを排除するとともに、エージェントの指差し方向が画像中の対象を正確に指すように配置を調整する。これにより、視聴者が注目すべき箇所を直感的に理解できるようになり、説明と動作の一貫性を維持しつつ、映像としての見やすさを確保する。

以上の結果、エージェントは「何を」「どこで」「どのように」提示すべきかを推定し、視聴者の注意を効果的に誘導できる。次節では、上記 (1) 画像切り出しモジュールの詳細について説明する。

## 4. 画像の切り出しの方法

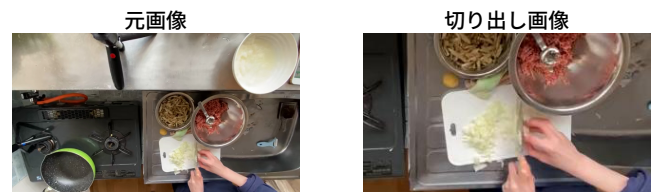
本研究の提案手法により調理画像を自動切り出した結果を図 3 に示す。提案手法では、調理シーンの画像と説明発話（タマネギを微塵切りにします。）を入力として、説明発話と最も関連する画像領域を自動的に抽出し、当該領域を切り出す。これにより、レシピの各工程における視覚的な焦点を明確化し、視聴者の理解を支援できる。

本システムは Python で実装されており、図 4 に示すように 3 つの主要ステップから構成される。本節以降では、各主要ステップの詳細について説明し、続いて、座標検証やアスペクト比調整などの補正処理について述べる。

### 4.1 GPT-4.1 を活用した画像切り出し

#### (1) 画像データのエンコード (図 4 上段)

本システムでは、入力された調理画像を OpenAI API に送信する前処理として、画像を Base64 形式にエンコードする。OpenAI API はバイナリ形式の画像データを直接受



説明発話：「タマネギを微塵切りにします。」

図 3: 説明発話に基づいた画像の切り出し結果の一例

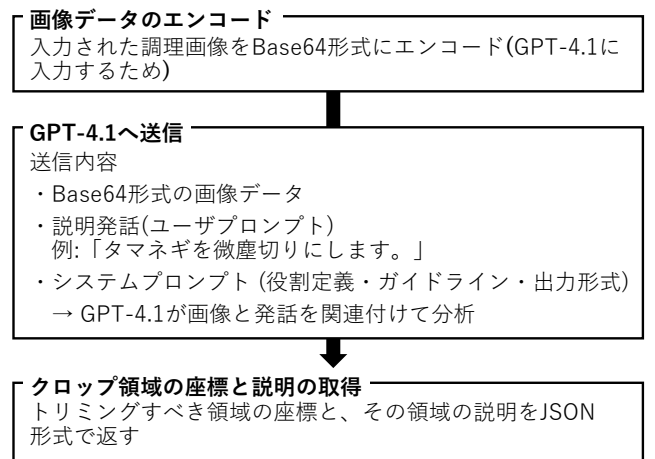


図 4: 提案手法の全体フロー（主要 3 ステップ）

け付けないため、この変換が必須となる。Base64 形式にエンコードすることにより、画像データをテキスト文字列として扱うことができる。

#### (2) GPT-4.1 へ送信 (図 4 中段)

Base64 形式にエンコードされた画像データは、ユーザが入力したレシピの説明発話（例：「タマネギを微塵切りにします。」）とともに GPT-4.1 モデルへ送信される。ここで GPT-4.1 には、COM-Kitchens に付与されたテキストを加工した説明発話を入力し、これに加えてシステムプロンプト（役割定義・ガイドライン・出力形式指定）を組み合わ

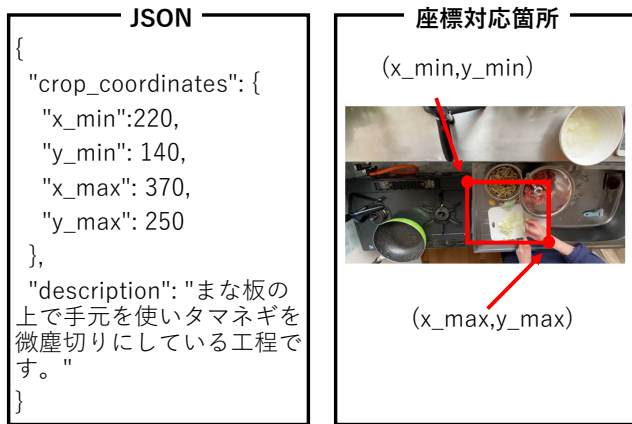


図 5: json 出力例と対応座標

せて送信する（システムプロンプトの設計方針については、4.3 節で詳述する）。説明発話はいくまで調理指示であり、「どの要素に注目すべきか」を直接含まないが、システムプロンプトによりモデルは発話内容と画像を関連付け、レシピに最も関連する領域、この場合は「包丁」や「まな板上のタマネギ」が含まれる領域の座標を出力する。

### (3) クロップ領域の座標と説明の取得（図 4 下段）

GPT-4.1 は、入力された画像と説明発話を受け取り、トリミングすべき領域の座標と、その領域が何であるかの説明を JSON 形式で返す。座標は、{"x\_min": 左上 X 座標, "y\_min": 左上 Y 座標, "x\_max": 右下 X 座標, "y\_max": 右下 Y 座標} の形式で提供される。また、"description" フィールドにはクロップ領域内で行われている調理工程や使用されている道具の概要が 1 ～ 2 文程度で記述されている。図 5 に、「タマネギを微塵切りにします。」という説明に対して返された実際の出力例を示す。左側には生成された JSON を、右側には元画像にクロップ矩形を重ねた結果を示している。

## 4.2 画像切り出し領域の補正

前述の GPT-4.1 を活用した切り出し手法により、注目すべき領域の抽出がある程度の精度で実現できるが、さらに提案手法では、抽出領域をエージェントの隣に表示する画像としてより安定的に扱うために以下の追加処理を行う。

### (1) 座標のチェックと調整

画像の端に近い領域が選択されると、後述の 16:9 アスペクト比調整処理により拡張された領域が画像境界を超える可能性がある。このため、システムは初期座標をチェックした上で、16:9 調整後の境界チェックという多段階での座標検証を行い、各段階で座標が負の値になっていないか、画像の幅や高さを超えていないかを確認し、必要に応じて座標を画像の端の点に制限する。

### (2) アスペクト比の調整（16:9 に統一）

見やすく統一感のある表示を実現するため、トリミング

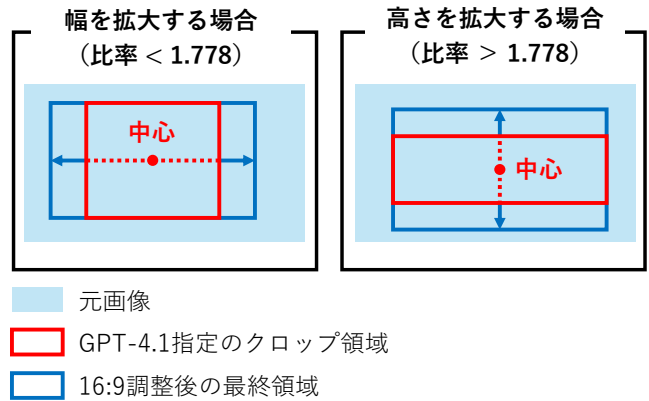


図 6: アスペクト比調整の模式図（左：幅を拡大／右：高さを拡大）

領域のアスペクト比を 16:9 に統一する。具体的には現在のクロップ領域の比率を計算し、目標比率 ( $16/9 \approx 1.78$ ) と比較する。現在の比率が目標比率より小さい場合は幅を拡大し、大きい場合は高さを拡大する（図 6）。調整は元のクロップ領域の中心を保ちつつ、左右または上下に均等に拡張することで実現される。画像の境界に達した場合は、境界内に収まるよう座標を調整し、必要に応じてピクセル単位での微調整を行う。この調整により、すべての切り出し画像の縦横比が統一された比率となる。

ただし、GPT-4.1 が指定した位置が画像の端に近い領域である場合、前述した「座標チェックと調整」により切り出し位置を画像の端に調整する場合がある。その際、中心位置が大きくずれ、不適切な結果となった画像については、アスペクト比の調整を行わない。

## 4.3 画像切り出しにおけるシステムプロンプトの設計

提案手法では、GPT-4.1 が一貫した出力を得るために、詳細なシステムプロンプトを設計している。本研究で実際に使用したシステムプロンプトの全文を付録 A.1 に掲載する。本プロンプトは以下の要素で構成される：

- (1) **役割定義**：料理画像解析の専門家としての位置づけ
- (2) **タスク定義**：レシピ指示文に対応する画像切り出しを行う際の座標の出力
- (3) **具体例の提示**：5 種類の調理動作パターン（「にんじんを切ります」「肉を炒めます」等）に対する探索対象の明示
- (4) **指示に従う際のガイドライン**：
  - 食材・調味料・手の動作を中央配置
  - 使用中/直前の調理器具のみ含める
  - 関係ない器具の除外
  - 抽象的動作時は道具と手を優先
  - 明確なシーンがない場合の代替基準
  - 複数候補時の優先順位
- (5) **技術的制約**：画像サイズ情報の提供と座標範囲の条件

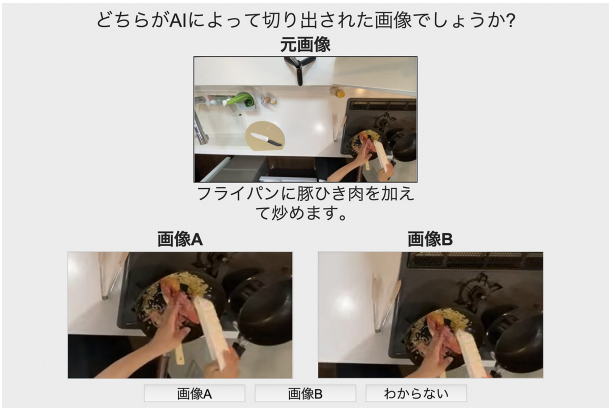


図 7: 評価の際に使用した画面

(6) 出力形式: JSON 形式での座標と説明文の指定

これらの構成により、調理に対するエージェントの説明発話が表す箇所を適切に切り出すことができるようになったと考えている。評価結果については次節で述べる。

5. 画像切り出しの評価実験

5.1 評価方法

提案手法による自動切り出し画像の妥当性を検証するため、人間による画像切り出しと比較し、切り出し精度を評価した。図 7 に評価に使用した画面を示す。図 7 に示す通り、画面上部に元画像と切り出し対象の説明発話を、その下に 2 枚の画像（「提案手法による自動切り出し」と「人間による切り出し」）を提示し、どちらが提案手法による自動切り出しかを知らされない状態で、提案手法による画像を回答させた。被験者は、「画像 A」、「画像 B」、「わからない」の 3 つの選択肢からひとつを選択する。各被験者あたり 5 問回答し、画像 A と画像 B のどちらに提案手法による自動切り出し画像が表示されるかはランダムとした。

被験者は、2025 年 7 月に行われた「愛知工業大学オープンキャンパス」におけるクイズ形式の展示「AI が切り抜いた画像を当ててみよう！」に参加した一般の方である。全体で 120 名が評価に参加し、各画像ペアに対して 5 名の被験者の回答を得た（計 600 件）。

5.2 評価データの作成

図 8 に示すように、評価対象となる各設問は、「入力画像」と「説明発話」の組で作成される。説明発話は調理の手順を表す文であり、人間と提案手法のどちらもこの説明発話と画像をもとに切り出しを行った。

評価データの作成手順は以下の通りである：

- (1) 人間と GPT-4.1 は共通の入力画像と説明発話に基づき、画像を切り出す。
- (2) 切り出された画像の中心位置を保持した上で、縦横比を 16:9 に統一する（中心が著しくずれる画像に関しては縦横比の補正は行わない）。



図 8: 評価画像の作成例

表 1: 正解率の集計結果

指標	値
総回答数	600 件 (120 問 × 各 5 件)
正解数 (提案手法を選択)	311 件
正解率	51.8%

- (3) 人間切り出し画像と提案手法による自動切り出し画像をペアとして組み合わせ、120 組の画像ペアを作成した。

なお、人間による切り出しは、愛知工業大学情報科学科 3 年生 12 名により実施し、1 人 10 枚ずつ担当。計 120 枚の切り出し画像を作成した。

5.3 評価結果

5.1 節で集めた被験者の回答を、「正解 (提案手法による自動切り出し画像を選択)」と「不正解 (人間による切り出し画像を選択または「わからない」を選択)」の 2 種類に分類し、正解率を提案手法による自動切り出し画像の自然さの指標として集計した。表 1 の「総回答数」は、120 問の評価データに対し、それぞれ 5 名の被験者から得られた回答の合計数 (120 問 × 各 5 件) を示している。「正解数 (提案手法を選択)」は、被験者が正しく提案手法による自動切り出し画像を選択した件数の合計である。「正解率」は、全回答数に対する正解数の割合を示す。

仮に提案手法による自動切り出しが人間の切り出しに劣っている場合、多くの被験者は 2 枚の画像を比較して容易に提案手法による自動切り出し画像を見分けられ、正解率は 100% に近づくと考えられる。しかし、今回の全体正解率は 51.8% であり、これはランダムに選択した場合の正解率 50% と大きな差がない。この結果は、被験者が提案手法による自動切り出し画像と人間による切り出し画像を明確に区別することが困難であったことを示しており、提案手法による画像切り出しが人間による切り出しと同程度の品質であることを示唆していると考えられる。

6. 事例分析

前節で述べた通り、提案手法による画像切り出しは人間による切り出しと同程度の品質であったが、一部適切でない切り出しもあった。本節では、被験者 5 名全員が提案手法による自動切り出し画像を正しく選択した画像（切り出



(a) 元画像

説明発話：「フライパンにたねを入れて焼きます。」



(b) 提案手法による自動切り出し画像



(c) 人間による切り出し画像

図 9: フライパンにハンバーグのたねを入れている画像の切り出し結果

し範囲が適切でなかったため、被験者に自動切り出しであることが見破られてしまった画像)に着目し、提案手法の代表的な誤りについて分析する。

#### 事例 1：フライパンにハンバーグのたねを入れている画像

図 9a は切り出し前の元画像で、フライパンにハンバーグのたねを入れている場面を示す。図 9b の提案手法による自動切り出し画像では、以下の特徴が見られる：

- **対象物の一部が欠けている**：フライパンや手の一部が画像外にはみ出しているため、動作全体の把握が困難である。
- **説明発話との関連性の弱さ**：画像上部には説明と直接関連しない背景情報が含まれており、行為の視覚的理解を妨げている。
- **構図の不自然さ**：被写体がやや中央からずれており、注目すべき部分が直感的にわかりにくい。

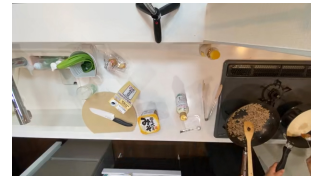
一方、図 9c の人間による切り出し画像では、フライパンや手元全体を明瞭に収め、説明発話の行為内容を直感的に理解できる構図となっている。

#### 事例 2：味噌を鍋に入れている画像

図 10a は切り出し前の元画像で、鍋に味噌を入れている場面を示す。図 10b の提案手法による自動切り出し画像では、以下の特徴が見られる：

- **説明発話との関連性の弱さ**：キッチンに置かれた味噌のパッケージが中心に切り出されており、説明発話の行為（味噌を鍋に入れる）との関連が薄い。
- **構図の不自然さ**：被写体の位置が端に寄っており、注目すべき行為が直感的に把握しにくい。

一方、図 10c の人間による切り出し画像では、鍋に味噌を加える動作が含まれており、説明発話が表示動作の全体像を捉えている。このことから、「味噌」などの材料と「鍋



(a) 元画像

説明発話：「味噌を鍋に入れます。」



(b) 提案手法による自動切り出し画像



(c) 人間による切り出し画像

図 10: 味噌を鍋に入れている画像の切り出し結果

に入れる」のような全体の動作の流れを捉える必要がある画像切り出しについては、提案手法は改善の余地があることが示唆された。

## 7. 結論

本研究では、調理動画における視聴者理解を支援することを目的として、エージェントによるジェスチャ付きインタフェースの構築に取り組んでいる。本稿では、説明発話の内容に合う部分を静止画像から切り出す手法を提案した。特に、視覚的に重要な領域を自然な構図で抽出するため、切り出し画像に含まれるべき物体に関する知識を GPT-4.1 モデルにプロンプトとして与えることで、意味的に妥当な切り出しを実現した。また、エージェントの隣に表示する画像としてより統一的な画像となるよう、画像の切り出し位置やサイズの補正を行う手法を提案した。

120 名の一般被験者の評価の結果、提案手法による自動切り出し結果は人間による切り出しと同等の精度であることが分かった。今後は、以下の三点に取り組む予定である。

- **ジェスチャの自動選択機構の構築**：発話の意味に応じて「切る」「混ぜる」などの動作を自動的に割り当てるため、動詞分類や形態素解析を活用したモーション選択アルゴリズムを開発する。
- **エージェントの動的配置の実装**：画像や発話内容に応じてエージェントの位置を動的に決定する処理を導入し、被写体との干渉や視認性を考慮した自然なレイアウト生成を実現する。
- **主観・客観の両面からの有効性評価**：開発したシステムが視聴者の理解促進に寄与するかを明らかにするため、アンケート調査などの主観的評価と、理解度テストによる客観的評価を併用して検証を行う。

以上の取り組みを通じ、対話型エージェントの視覚的補助の精度を高めることで、調理動画における学習支援およ

び安全性向上に寄与するシステムの構築を目指す。

## 参考文献

- [1] Surgenor, D., Hollywood, L., Furey, S., Lavelle, F., McGowan, L., Spence, M., Raats, M., McCloat, A., Mooney, E., Caraher, M. and Dean, M.: The impact of video technology on learning: A cooking skills experiment, *Appetite*, Vol. 114, pp. 306–312 (2017).
- [2] Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A. and Murphy, K.: What's Cookin'? Interpreting Cooking Videos using Text, Speech and Vision, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, pp. 143–152 (2015).
- [3] Doman, K., Kuai, C. Y., Takahashi, T., Ide, I. and Murase, H.: Video CookIng: Towards the Synthesis of Multimedia Cooking Recipes, *Advances in Multimedia Modeling*, Springer Berlin Heidelberg, pp. 135–145 (2011).
- [4] Nishimura, T., Hashimoto, A., Ushiku, Y., Kameko, H. and Mori, S.: Recipe Generation from Unsegmented Cooking Videos, *ACM Trans. Multimedia Comput. Commun. Appl.* (2024).
- [5] Fujii, T., Orihara, R., Sei, Y., Tahara, Y. and Ohsuga, A.: Generating Cooking Recipes from Cooking Videos Using Deep Learning Considering Previous Process with Video Encoding, *Proceedings of the 3rd International Conference on Applications of Intelligent Systems*, APIS 2020, Association for Computing Machinery (2020).
- [6] Maeda, K., Hirasawa, T., Hashimoto, A., Harashima, J., Rybicki, L., Fukasawa, Y. and Ushiku, Y.: COM Kitchens: An Unedited Overhead-view Video Dataset as a Vision-Language Benchmark, *Proceedings of the European Conference on Computer Vision* (2024).
- [7] 井上昂治, ラーラーディベッシュ, 山本賢太, 中村 静, 高梨克也, 河原達也: アンドロイド ERICA の傾聴対話システム—人間による傾聴との比較評価—, *人工知能学会論文誌*, Vol. 36, No. 5, pp. H-L51.1–12 (2021).
- [8] Tokuda, K. et al.: MMDAgent: A fully open-source toolkit for voice interaction systems, *ICASSP* (2013).
- [9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision, *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 139, PMLR, pp. 8748–8763 (2021).
- [10] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I. et al.: GPT-4 Technical Report, arXiv:2303.08774 (2024).

## 付 録

### A.1 システムプロンプト全文

本研究で利用したシステムプロンプトの全文を、図 A-1 に示す。

あなたは料理画像解析の専門家です。

ユーザーから送られてきた料理画像に対して、レシピ指示に関連する部分をクロッピングするための座標を提供してください。

具体例:

- 「にんじんを切ります」→ 画像内で切られている、または切る準備ができてにんじんや包丁、まな板などの調理器具が収まるように探す
- 「肉を炒めます」→ フライパンで炒められている肉や、これから炒める予定の肉を探す
- 「ボウルで混ぜます」→ 材料が入ったボウルを探す
- 「油をひきます」や「フライパンに油をひきます」→ 油を注いでいる手元、油の容器、フライパンの表面、あるいはその直前の準備動作が見られる場所を探す
- 「醤油を入れます」→ 醤油の容器やフライパン、鍋を探す

指示に従う際のガイドライン:

1. 食材や調味料、または「手の動作」が画像の中央付近に来るような範囲を優先してクロップしてください。
2. 調理器具は、現在使用されている・もしくは使用直前であるもののみを含めてください。
3. 関連する食材と関係のない調理器具は可能な限り含めないでください。
4. 動作が抽象的で画像から明確でない場合(例:「油を引く」)は、動作に使われる道具(油のボトル、フライパン)と、手が関与している部分を優先してクロップしてください。
5. 明確な調理シーンが存在しない場合は、レシピ指示に最も関係する道具や食材を中央に配置した状態でクロップ範囲を決定してください。
6. どうしても複数の候補がある場合は、最も関連性が高く、かつ視覚的に分かりやすい部分を優先してください。

画像のサイズは幅{img\_width}ピクセル、高さ{img\_height}ピクセルです。  
このサイズ内で有効な座標を返してください。

あなたの出力は必ず以下のJSON形式に厳密に従ってください:

```
{{
  "crop_coordinates": {{
    "x_min": 整数値,
    "y_min": 整数値,
    "x_max": 整数値,
    "y_max": 整数値
  }},
  "description": "このクロップ画像はどんな料理道具を用いてどのような料理工程を行なっているか一言で書いてください。"
}}
```

説明と注意点:

- x\_min, y\_min は左上の座標、x\_max, y\_max は右下の座標です
- 座標は元の画像のピクセル単位で整数値で指定してください
- 必ず有効な座標を返してください (x\_min < x\_max かつ y\_min < y\_max)
- 必ず画像範囲内の座標を指定してください (0 ≤ x\_min < x\_max ≤ {img\_width} かつ 0 ≤ y\_min < y\_max ≤ {img\_height})
- JSONフォーマット以外のテキストや説明は一切含めないでください
- 必ず有効なJSONを出力してください
- コードブロック記号(``)は含めないでください

図 A-1: 本研究で利用したシステムプロンプト