

大規模言語モデルからの知識抽出に基づく 画像からのスクリプト予測の検討

八木 拓真¹ 西村 太一² 清丸 寛一² 唐井 希²

¹ 東京大学 生産技術研究所 ² 京都大学大学院 情報学研究科
 tyagi@iis.u-tokyo.ac.jp kiyomaru@i.kyoto-u.ac.jp
 {nishimura.taichi.43x, karai.nozomu.77c}@st.kyoto-u.ac.jp

概要

画像からそれが示す場面で行われるスクリプト (典型的なイベント系列) を予測するタスクを提案する. 画像が示す状況に対応したスクリプトを想起することにより, エージェントが状況に応じた適切な行動理解や支援を行うことが期待できる. 本稿では, 画像からのスクリプト予測の第一歩として少数の画像にアノテーションを付与し, その要件と課題の整理を行う. また, 学習済みの視覚言語モデルと大規模言語モデルを組み合わせることによる画像からのスクリプト予測の可能性を検討する.

1 はじめに

人と AI が共存する社会の実現にあたり, ロボットなどのエージェントが場の状況に応じて適切に振る舞うことは重要である. 人はある場面に遭遇した際, その場面に対応したスクリプト (典型的なイベント系列) [1] を想起することで現在の状況をスムーズに理解し, 適切に振る舞える. 例えば, レストランでは一般に席に案内され, 料理を注文し食事後会計をすることが期待できる. 従来スクリプトは文章の論理構造を補完するものとして言語に閉じて扱われてきた [2]. しかし, 家庭用ロボットなどへの応用を考えれば, 状況が詳細に文章で与えられなくとも, 視覚などの外界の入力から直接スクリプトを想起し行動に移せることが望ましい.

本稿では, 画像からその場面に対応するスクリプト [1] を予測する視覚スクリプトモデルの検討を行う (図 1). スクリプト知識は宣言的知識と比較して構造があいまいであり, 「レストランで食事を取る」といった典型的な状況であっても場面により多様なバリエーションを持つ. そのため, モデルの訓練のために幅広い場面对するスクリプト知識を人手で



図 1 視覚スクリプト予測の問題設定.

網羅的に収集するのは現実的でない. そこで, 本研究では大規模言語モデル [3] (Large Language Model; LLM) から知識を抽出する方法を検討する. LLM は様々な常識推論タスクに回答できることが報告されており, 典型的な状況におけるスクリプト知識をかなりの程度獲得していると思われる. 画像中の状況をテキストで説明する視覚言語モデルと大規模言語モデルを組み合わせることにより画像とスクリプトを結び付けられれば, スケーラブルな形で訓練用のデータを収集できることが期待できる.

本稿では, 画像からのスクリプト予測の初期検討として少数の画像にアノテーションを付与し, 視覚スクリプトモデルの実現のための課題と要件を整理する. また, 学習済みの視覚言語モデルと大規模言語モデルを組み合わせることによる素朴なベースラインモデルを構築し, (i) 画像およびゴールからのスクリプト予測 (ii) 画像からのゴール予測の 2 種類のタスクに関する予測結果の評価を行う.

2 関連研究

2.1 大規模言語モデル

近年、個別のタスクに対する専用モデルに代わり、数百～数千億のパラメータを持ち、大規模コーパス上で事前学習された汎用の LLM [4, 5, 3, 6] の有効性が実証されている。LLM は様々な常識推論タスクにファインチューニングなしで回答できるうえ、モデルに与える入力文（プロンプト）を工夫することによりさらに性能を向上できることが報告されている [7] が、入出力共にテキストに閉じているため、単体では実世界に接地した推論は行えない。

また、LLM と同様、大量の動画像とテキストの組によって事前学習を行う視覚言語モデル（Visual Language Model; VLM）も注目されている [8, 9, 10]。VLM は画像に写る事物の説明や質問応答には長けているものの、画像に直接写らない文脈や手続きの認識はごく少数の初期検討 [11] があるのみで十分には検討されていない。

2.2 イベント系列の予測

家庭用ロボットなどでは、対話を通じて人の指示を理解し、自身の行動を実行可能な形で推論する必要がある。そこで、言語指示からのスクリプトの予測 [12, 2] やユーザの要求に応じた行動のプランニング [13, 14, 15] 等の研究が行われており、LLM の利用も検討されている [15, 16]。これらは限定的なシナリオで具体的な指示が与えられることを前提としているが、本研究では様々な場面における画像からスクリプト知識を想起することを目的とする。

3 視覚スクリプト

本節では、画像からその場面の観察者または登場人物から見たスクリプトを予測することを**視覚スクリプト**と定義し、その要件と課題を整理する。

3.1 アノテーション

画像はスクリプトに関して何を伝えるだろうか？ 画像が示す**場所**はそこで何をすべきかの第一の手掛かりになる。画像中の**物体**やその見た目もその場の属性を示唆し、例えば高級レストランと家庭的なレストランとではその備品や内装は大きく異なり、生起するスクリプトも異なる。画像中に登場する人物の**役割**はその人物の行動を規定する。一方、

表 1 アノテーション項目の一覧。

場所	画像が示す場所名
ゴール	当該場面で典型的に設定される目標
主役	スクリプトの実施主体の属性
脇役	スクリプト中に登場する他の人物や主体
物体	スクリプト中に登場する物体と資産
スクリプト	ゴールを達成する典型的なイベント系列

表 2 アノテーションの統計（カッコ内は標準偏差）。

画像枚数	40
アノテーション数	125
平均系列長	5.6 (±1.8)
平均脇役数	1.3 (±0.7)
平均物体数	2.5 (±1.6)

スクリプトを行う主体がどのような**役割**を担うか、何を**ゴール**として行動するかといった内的な情報は通常画像には含まれず、実応用では観察者（人、ロボットなど）の経歴や過去の行動に基づきそれらを補完する必要がある。オクルージョンのため、不完全な観測から見えない部分を補完する必要もある。

Schank & Abelson [1] は、登場人物・物体や完了条件といったスクリプトの構成要素を提案しており、本研究においてもそれらが有用と考えた。そこで彼らの定義を踏襲し、表 1 に示す項目について画像に対応するスクリプトのアノテーションを行った。題材としては場所認識用データセットである Places365 [17] データセットを用い、大きく (i) 購買 (ii) 室内 (iii) 交通 の 3 つの場面を中心に 26 カテゴリ 40 枚を選定した。各画像について、3 人のアノテータが 1 枚あたり 1 つ以上のスクリプトを付与した。ただし、場所に関しては原則 Places365 に既に付与されている名前を採用した。スクリプトの記述にあたっては、アノテータが**画像が示す状況において誰**（行為者）が何（ゴール）を達成できるかを類推したうえで、それを達成するイベント列を記述した。

3.2 アノテーションの分析

表 2 に統計を示す。スクリプトの系列長は概ね 5-6 程度であり、大半の例において主役以外の脇役が登場した。登場物体は最大 7 つ程度であり、ゴール達成のための不可欠なイベントを簡潔に記した例が多くを占めた。また、表 3 に図 1 へのアノテーション例を示す。この例では、プレゼンの準備をする/カンファレンスでの発表の準備/会議へ参加/発表を聞くの 4 つのユニークなゴールが付与され、他の画像についても主役/主役の動機/ゴールの粒度の違いによる多様性が見られた。一方、購買を伴う場所

表3 図1の画像に対するアノテーション例.

場所	Conference center (Places データセットを踏襲)	
主役	Presenter	Participant
ゴール	Prepare for the conference	Attend a meeting
脇役	Conference staff	Host, speaker
物体	Chair, screen, microphone	Microphone, chair, screen
スクリプト	Head to the room where the presentation is to be made	Sit on a seat
	Tell the conference staff that you are a presenter and would like to check the equipment	Listen to the speaker's presentation
	Confirm that the PC can be connected to the projector	Raise your hand if you have questions or comments
	Confirm that the PC's audio is coming out of the room's speakers	Receive a nearby microphone
	Tell the conference staff that you have completed the check	Ask a question or express an opinion
	Thank the conference staff	Listen to the speaker's response

の画像ではゴールが取れんする傾向にあった。

難しさとしては (i) 画像に見える事物とルールや慣習との関連の非自明さ (ii) 画像に登場しない役割や物体の補完があった。例えば、店舗では購買行動が行われるが、売り物の見た目自体にその行動が結び付く必然性はなく、物体の組み合わせとしてどのような場所かを知識により結び付ける必要がある。また、物を買う場面では通常会計係、伝票やレジスターが登場するが、それらが明示的に写ることはまれであり、多くの場合それらを補完する必要がある。見える範囲の即物的に記述は不十分であり、見える範囲を超えた推論の必要性が示唆された。

4 実験

前節の議論で、画像中の情報は生起するスクリプトを規定する一方その関係は間接的であることが示された。そのため視覚スクリプト生成においても LLM に含まれる豊富な知識を活用することが望ましい。画像情報を反映したスクリプトを実際に LLM を用いて生成できるのか？どの要素が正しい生成に寄与するのか？上記の初期検討として、VLM と LLM を素朴に組み合わせたベースモデルを提案し2種類のタスクによる評価を行った。

4.1 画像とゴールからのスクリプト予測

まず、画像とゴールなどの付加情報から LLM を通じてスクリプトを生成できるかを検証した。

モデル LLM はテキストのみを受け付けるため、画像情報を LLM に入力可能な形に変換する必要がある。そこで、最新の画像説明文生成モデル [9] により説明文を生成し LLM の入力として用いるモデルを提案する。図2に模式図を示す。まず、入力画像の見える範囲での説明文を画像説明文生成モデル [9] を用いて生成し、画像をテキスト情報に変換

表4 スクリプト予測の人手評価結果 (%)。

言語モデル	合理性	適切さ	
		ゴール	画像
GPT-3	98.5	97.7	96.7
OPT-175B	82.3	79.2	83.1
プロンプトに関する ablation (GPT-3)			
場所名のみ	93.8	-	81.3
ゴールのみ	100.0	64.7	53.1
説明文のみ	100.0	-	51.3

する。続いて、前節で付与した情報のうち場所、主役およびゴールを説明文と共に LLM のためのプロンプト (入力文) として構成し、それを LLM を通すことによってスクリプトを得る。入力方法には LLM で一般に用いられる in-context learning [3] を採用し、1 サンプルに関する見本を与えたのちスクリプトをイベント列として出力させた。

実装と比較条件 画像からの説明文生成には最新のモデルとして BLIP-Large [9] を使用し、LLM は GPT-3 (text-davinci-003) [3] と OPT (OPT-175B) [6] の2種類を比較した。また、入力情報の種類による違いを見るため GPT-3 では (a) 場所名のみ (b) ゴールのみ (c) 説明文のみの3種類の ablation を行った。具体的な予測例は付録表6を参照のこと。

評価指標 一場面に対応するスクリプトは多様であり、自動評価は難しい。そこで前項のアノテーションのうち類似のゴールを除いた40画像65サンプルに対しスクリプトを生成し、著者らのうち2名の人手評価により合理性 (スクリプトとして妥当か)、適切さ (ゴール・画像に照らして適切か) の3指標を二値で判定しその平均を取った。

結果 表4に結果を示す。画像説明文およびゴールなどの付加情報を与えた LLM はゴール・画像の両者に対して高い適切さを報告し、GPT-3 を用い

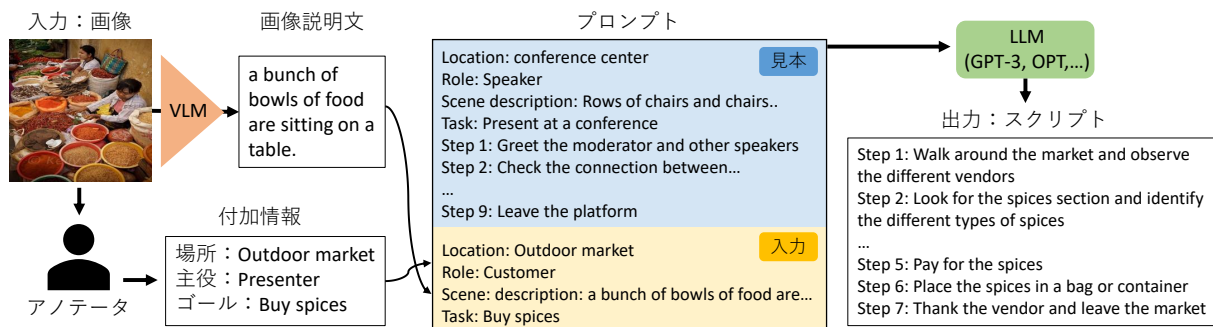


図2 VLMとLLMの結合による画像からのスクリプト予測の模式図. 本稿ではVLMを通して生成した説明文および人手で付与した付加情報を組み合わせてプロンプトを生成し、それをLLMに通してスクリプトを予測する。

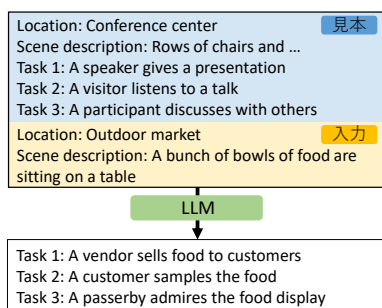


図3 場所名および説明文からのゴール予測.

表5 ゴール予測の人手評価結果 (%)

条件	合理性	適切さ	
		場所	画像
場所名のみ	97.1	95.0	92.5
説明文のみ	77.9	56.7	51.7
場所名+説明文	92.5	86.3	83.3

た場合にはほぼ望ましいスクリプトを生成した。OPT-175Bも概ね同様の挙動を示したものの不自然な動作を繰り返す、扱う物体や行動を取り違えるといった誤りを示した。多くのスクリプトは場所に依存するため、場所名のみを与えた場合も81.3%と画像に対する高い適切さを示した。しかしながら、路上の市場の画像に対して高級レストランと同様のやりとりを提示するなど、場所名のみからは読み取れない文脈において不適切なスクリプトを生成した。ゴールと説明文をそれぞれ単体で与えた場合も同様で、前者では店舗で会計しない、後者では目に見える物体を用いたランダムな行動を行うなどのエラーが見られ、場所情報に画像から得られる情報を加えることでよい生成ができることを確認した。

4.2 画像説明文からのゴール予測

前項はスクリプトの生成に注目し主役およびゴールを既知としたが、それらも自動生成できれば場所

認識含め全処理を自動化できる。そこで類似の枠組みで説明文と場所からのゴール予測を評価した。

モデルと評価 前項のモデルと同様の枠組みを採用した。図3にLLM部のプロンプトと出力の例を示す。今回は3つのゴールを出力するようなプロンプトを与えた。LLMはGPT-3を使用し、(a)場所名のみ (b)説明文のみ (c)場所名+説明文の3条件を比較した。評価指標としては前項と同様ゴールとしての適切さおよび場所・画像に対する適切さを項目ごとに二値で判定し、その平均を取った。

結果 表5に結果を示す。我々の予想に反し、場所名のみを与えた場合に高い適切さを示した。スクリプト中のイベントより高い抽象度でのゴールは画像の即物的な説明より場所による影響が大きかったからと思われる。失敗例としては、惣菜店でウェイターが料理を給仕するなどの不適当な登場人物や、犬がおもちゃで遊ぶといったスクリプトの条件を満たさない場合等があった。しかしながら、多くのサンプルにおいて適切な役割とゴールを生成できたことから、指示文や画像説明文の生成方法の工夫次第でより性能を伸ばせると考えられる。

5 結論と将来展望

画像からその場面に対応するスクリプトを予測する問題に初めて取り組み、アノテーションの付与および画像からのスクリプト・ゴール認識の初期検討を行った。評価を通じてVLMとLLMの素朴な結合モデルが画像に対応したスクリプトを概ね正しく生成できることを確認し、LLMからの知識抽出の可能性を示した。今後は(i)より大規模の評価用コーパスの作成(ii)洗練されたプロンプティングによる制御可能なスクリプト生成(iii)陽な説明文を介さない、VLMとLLMを融合した視覚スクリプト予測モデルの実現を目指す。

謝辞

本研究は JST ACT-X JPMJAX22AJ の支援を受けたものです。

参考文献

- [1] Roger C Schank and Robert P Abelson. **Scripts, plans, goals, and understanding: An inquiry into human knowledge structures**. Psychology Press, 1977.
- [2] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. proScript: Partially ordered scripts generation. In **Proc. Findings of EMNLP**, pp. 2138–2149, 2021.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In **Proc. NeurIPS**, pp. 1877–1901, 2020.
- [4] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proc. NAACL-HLT**, pp. 4171–4186, 2019.
- [5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **JMLR**, Vol. 21, No. 140, pp. 1–67, 2020.
- [6] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. **arXiv preprint arXiv:2205.01068**, 2022.
- [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In **Proc. NeurIPS**, 2022.
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **Proc. ICML**, pp. 8748–8763, 2021.
- [9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In **Proc. ICML**, pp. 12888–12900, 2022.
- [10] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. **TMLR**, 2022.
- [11] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In **Proc. CVPR**, pp. 6720–6731, 2019.
- [12] Qing Lyu, Li Zhang, and Chris Callison-Burch. Goal-oriented script construction. In **Proc. INLG**, pp. 184–200, 2021.
- [13] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In **Proc. CoRL**, 2022.
- [14] 田中翔平, 湯口彰重, 河野誠也, 中村哲, 吉野幸一郎. 気の利いた家庭内ロボット開発のための曖昧なユーザ要求と周囲の状況の収集. 情報処理学会研究報告, Vol. 2022-NL-253, No. 5, pp. 1–7, 2022.
- [15] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In **Proc. ICML**, pp. 9118–9147, 2022.
- [16] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. ProgPrompt: Generating situated robot task plans using large language models. **arXiv preprint arXiv:2209.11302**, 2022.
- [17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. **TPAMI**, pp. 1452–1464, 2017.

表6 図2の画像に対するVLMとLLMを用いたスクリプトの生成例.

入力文	Location: conference center Role: Speaker Scene description: Rows of chairs and chairs in a room with chairs and a projection screen. Task: Present at a conference Step 1: Greet the moderator and other speakers Step 2: Check the connection between the computer and the screen and speaker Step 3: Listen to other speakers' presentations until the stage Step 4: At your turn, stand in front of the platform and connect your laptop. Step 5: The moderator introduces the biographies of the speakers Step 6: Give a presentation using a microphone Step 7: Answer questions from the listeners Step 8: Disconnect the laptop Step 9: Leave the platform
	Location: Outdoor market Role: Customer Scene: description: a bunch of bowls of food are sitting on a table Task: Buy spices
モデル	応答
GPT-3	Step 1: Walk around the market and observe the different vendors Step 2: Look for the spices section and identify the different types of spices Step 3: Ask the vendor about the quality and price of the spices Step 4: Select the spices you want to buy Step 5: Pay for the spices Step 6: Place the spices in a bag or container Step 7: Thank the vendor and leave the market
OPT-175B	Step 1: Look at the bowls of food Step 2: Choose a bowl of food Step 3: Look at the price tag Step 4: Decide whether to buy the bowl of food Step 5: Ask the seller for the price of the bowl of food Step 6: Pay the seller Step 7: Take the bowl of food
GPT-3 (場所名のみ)	Step 1: Arrive at the market and look around Step 2: Talk to the vendors and ask questions about their products Step 3: Select the items you want to buy Step 4: Bargain with the vendors to get the best price Step 5: Pay for the items Step 6: Place the items in a bag or basket Step 7: Thank the vendors for their help Step 8: Leave the market and enjoy your purchases
GPT-3 (説明文のみ)	Step 1: Gather the necessary utensils such as plates, forks, and spoons. Step 2: Place the bowls of food on the table. Step 3: Invite guests to the table. Step 4: Pass around the bowls of food and let each guest serve themselves. Step 5: Enjoy the meal together. Step 6: Clean up the table and put away the utensils.
GPT-3 (ゴールのみ)	Step 1: Research the types of spices you need Step 2: Make a list of the spices you need Step 3: Check the prices of the spices online Step 4: Compare prices from different stores Step 5: Choose the store with the best prices Step 6: Place an order for the spices online Step 7: Pay for the spices Step 8: Wait for the spices to be delivered Step 9: Enjoy your spices!

A VLMとLLMを用いたスクリプトの生成例

図2の画像に対してVLMとLLMを組み合わせることにより生成したスクリプト文の具体例を表6に示す。LLMがより望ましい出力を行うよう、一部の用語はプロンプトでは別の単語に置き換えられている場合がある。上段に入力文を、下段にモデルの応答例を示す。紙面の都合で応答の改行は省略した。GPT-3はゴール(buy spices)を達成するための行動を適度に列挙し、そのための手順を適切に生成している。OPT-175Bはよりシーンの説明(ボウルに入った食べ物)に注目したスクリプトを生成し、やや流暢さには欠けるが適切と言える。場所名(outdoor market)のみを与えた場合は市場での典型的な行動を反映しているが具体性には欠ける。説明文のみを与えた場合には目の前の食べ物を用いてゲストをもてなす等のゴールとは無関係のスクリプトが生成されている。ゴールのみを与えた場合は確かにスパイスを買うための行動が列挙されているが、今どのような場所にいるかが不明なためインターネット上での購入方法が書かれており、これは画像に照らし合わせて適当ではない。このように、場所・説明文・ゴールのいずれも画像の示す場面に照らし合わせて適当なスクリプトの生成のために必要な情報であることがわかる。