# 視覚情報を用いたタスク指向型対話における人間の応答に対する間違い指摘の検討

大島遼祐 [1]　品川政太朗 [2]　綱島秀樹 [3]　森島繁生 [4]
[1] 早稲田大学 先進理工学部　[2] 奈良先端科学技術大学院大学 先端科学技術研究科
[3] 早稲田大学 先進理工学研究科　[4] 早稲田大学 理工学術院総合研究所
ryosukeoshima@fuji.waseda.jp　sei.shinagawa@is.naist.jp
h.tsunashima@asagi.waseda.jp　shigeo@waseda.jp

## 概要

視覚を持つ対話システムが実世界応用される際，対話相手である人間のミスが原因でタスクが失敗する場合があり，システムには失敗を避けるために人間を支援することが求められる．そこで本研究では，視覚情報を用いたタスク指向型対話の一つである Guess What?!での人間を支援する方法として，人間の応答ミスを指摘するという問題設定を新たに考える．人間が応答に間違えたデータの収集・分析を行い，間違い指摘モデルの正確性向上への有効的な学習方法と入力の検証を行った．実験の結果から，人工的なデータセットによる事前学習と，応答時刻・質問タイプを加えるという二つが間違い指摘の正確性向上に部分的に寄与する事が分かった．[1)]

## 1　はじめに

近年，テキスト情報だけでなく，視覚情報も考慮してチャット形式の対話行うタスクである Visual Dialogue の研究が盛んである [1–4]．特に，Guess What?! [2]（図 1）は二人で協力して行うゲームとなっており，一方 (Oracle) が参照している画像内の物体 (以下，正解物体) を，もう一方 (Questioner) が質問応答を繰り返すことにより推測するという，タスク指向型対話になっている．

Guess What?!を始めとして，現状の多くの視覚情報を用いた対話タスクでは，一方の質問に対するもう一方の応答は，質問への応答として常に正しい事が前提になっている [4–7]．しかし，現実の対話では，Oracle 側を担当する人間やモデルが間違えた回答をしてしまう場合も考慮する必要がある．

そこで，本研究では Questioner を対話エージェントが，Oracle を人間が担当する場面を考え，人間がエージェントの質問に対する応答を間違えた場合に，エージェントがミスを指摘するといった新しい問題を提案する．既存研究 [8] でも本研究と同じく，対話相手の応答が間違える場合の重要性を指摘しているが，こちらは対話相手がタスクを失敗させようと非協力的にわざと不適切に応答している状況下で，非協力的であることに気づくという問題設定である．本研究は，対話相手が協力的にも関わらず応答に間違えてしまう状況を想定し，実際に間違いの応答を指摘するという点で異なる．

| Questioner | Oracle |
|---|---|
| 1. Is it a donut? | Yes |
| 2. Is it coated with green? | **Yes** |
| 3. A complete donut? | Yes |

**図 1**　Guess What?!ゲームの一例．正解物体は水色枠に囲われている．この例では，2 番目の応答が間違いである．

人間の応答ミスの指摘は，いつ指摘するかという視点から大きく 3 つに分けられると考えられる．

1. 対話中に人間の応答ミスに気づき指摘する．応答ミスによりタスクが失敗の方向に向かっていたが，エージェントがミスに気づき指摘する．

1）　本論文の内容は 2023 年の情報処理学会 第 85 回全国大会で発表した内容と同一の内容を一部含みます．

2. 物体の推測に失敗してしまった後に，人間の応答のミスに気づき指摘する．失敗で終わらせずそこから復帰して，物体の推測を成功させようというものである．
3. 物体の推測に失敗してしまった後に，人間から正解物体をエージェントが教えられ，それに基づいて人間の応答のミスに気づき指摘する．

本研究では，対話相手である人間の応答の間違いを指摘できるようになるための第一歩として，3 番の問題設定を考える．この問題設定は，最も容易であるが，次に全く同じようなタスク指向型対話を行った際に先ほどの教訓を活かして，自分の応答により注意を払うことができるようになり，人間の応答ミスによるタスク失敗の数を減らせるようになることが期待できると考える．

本研究では，人間の応答ミスの指摘モデルに望ましい性質とは何か検証するため，まず人間が応答にミスをしたデータの収集と分析を行い，続いてその分析に基づいて，収集したデータから得られる知見の有効性を実験により検証する．なお，間違い指摘の問題設定において学習データが少ないという問題があるため，少ないデータでミスの指摘をより可能にする人工的なデータセットによる事前学習方法を提案し，その有効性についても併せて検証する．

# 2 人間の応答ミスの収集と分析

Guess What?!データセット [2] には人間同士による対話ゲームの結果が収録されている (約 15 万対話，82 万個の質問応答ペア，6.6 万個の画像)．また，Questioner が正解物体の予測に成功したもの「status = Success」，失敗したもの「status = Failure」，途中で諦めたのもの「status = Incomplete」の 3 つのラベル付けされている．

## 2.1 応答ミスの含まれるデータの収集

本研究では，実際に人間が Guess What?!ゲームにおいて応答に間違えたデータを集めた (以下，**手動データセット**)．具体的には「status = Failure」のラベル付けがされたデータセットから 2,300 対話をランダムにサンプリングし，人間の応答にミスがないか手動でチェックすることで構築した．構築の際，正解物体の大きさが小さすぎるもの，同一の対話に応答ミスが 3 つ以上含まれるものは除外した．ミスが 3 つ以上のものを除外したのは，人間 (Oracle) が最初から適当に答えている可能性が高いと判断したためである．その結果，365 対話からなる応答に間違いが含まれる手動データセットを構築することができた．

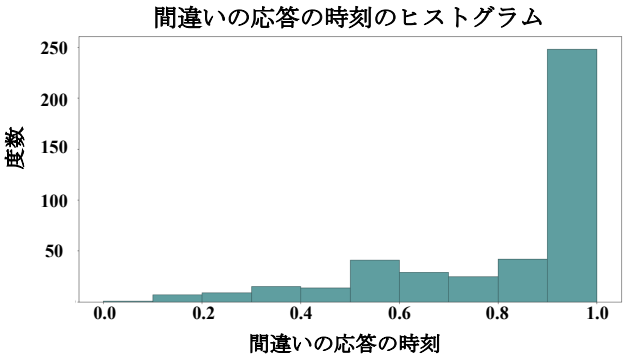


**図 2** 応答時刻と応答ミスの関係性のヒストグラム．横軸の数値は正規化されている．

**表 1** 時刻と応答ミスの関係

| | 最終時刻 | 最終時刻以外 |
|---|---|---|
| 応答ミス数 | 231 | 200 |

## 2.2 応答ミスの含まれるデータの分析

まず，本研究では人間の応答ミスの特徴を見るために，手動データセットに関して分析を行った．具体的にはデータ収集の際，時刻によって応答を間違える頻度が異なる印象を受けたため，応答時刻と応答ミスの関係性の分析（2.2.1 項）を行った．また [9, 10] において，Oralce モデルが質問タイプによって応答の正答率が異なることが言及されてるため，質問タイプと応答ミスの関係性（2.2.2 項）についても分析を行った．

### 2.2.1 応答時刻と応答ミスの関係性

図 2 に応答時刻と応答の間違いの関係性のグラフを示す．図 2 の横軸は，以下の計算に基づいて正規化されている．

$$\text{横軸} = \frac{\text{応答時刻}}{\text{対話の長さ}} \tag{1}$$

図 2 より，人間の応答ミスは対話の後半部分にいくにつれ，多くなっていることがわかる．また，表 1 に最終時刻に応答ミスをしている数と，最終時刻以外に応答ミスをしている数を表す．表 1 から，人間は最終時刻の応答を間違えやすい傾向にあることがわかる．図 2 と表 1 のような結果が得られたのは，対話が後半になるほど，細かく物体を識別し正解物体を当てるために，より複雑な質問（例：“Is it the man left side from the man wearing a red hat?”）が多くなるためであると考えられる．

### 2.2.2 質問タイプと応答ミスの関係性

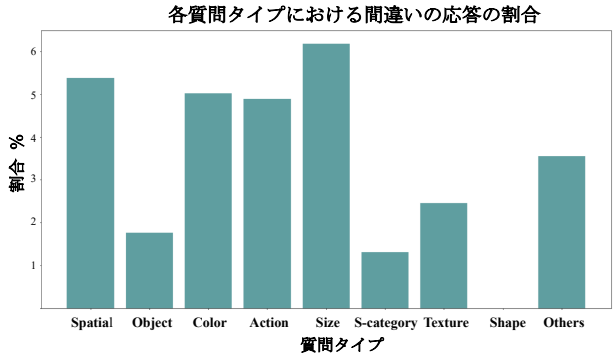


**図 3** 各質問タイプにおける間違いの割合．S-category は，Super-category タイプのことを表す．

図 3 に質問タイプと応答ミスの関係性のグラフを示す。質問タイプのラベル付けは，[9, 10] に倣い，キーワードマッチング方式[2)]で行った。質問はそれぞれ，**Object** (“is it a car?”)， **Spatial** (“on the right side half?”)，**Color** (“is it white?”)，**Action** (“are they wearing jeans?”)，**Size**(“a small one?”)，**Super-category** (“is the object an electronic?”)，**Texture** (“is it made of metal?”)，**Shape** (“is it a round container?”) ，**Others** の 9 種類に分類されている． ここで，括弧内は各質問タイプの例文である．図 3 からわかるように，人間は **Spatial**，**Color**，**Action**，**Size** のいずれかに分類される質問に対して，間違えやすいことがわかった．以上の分析から，Guess What?!における人間の応答は，応答時刻と質問タイプによって間違えやすい傾向が異なることがわかった．

## 3 実験

本研究では，２種類の実験を行った．1 つ目は，手動データセットの大きさが小規模である問題点に対し，提案する事前学習方法が有効か検証する実験（以下，実験 1）である．2 つ目は間違い指摘をより正確に行うには何の入力が重要かを検証する実験（以下，実験 2）である．2 章で得られた人間の応答の間違いの傾向が，どのように間違い指摘モデルの精度向上に寄与するか検証した．

**比較モデル**　本実験は，学習方法とモデルの入力を検証する実験のため，実験 1 と実験 2 における比較モデルとして，図 4 のような非常にシンプルなモデルを構築した．具体的には，Guess What?! [2] で提案された Oracle モデルに，間違い指摘のための分類ヘッドを追加し，モデルの出力に対し閾値を 0.5 として，応答が{間違っているか，合っているか}を 2 値分類[3)]する．

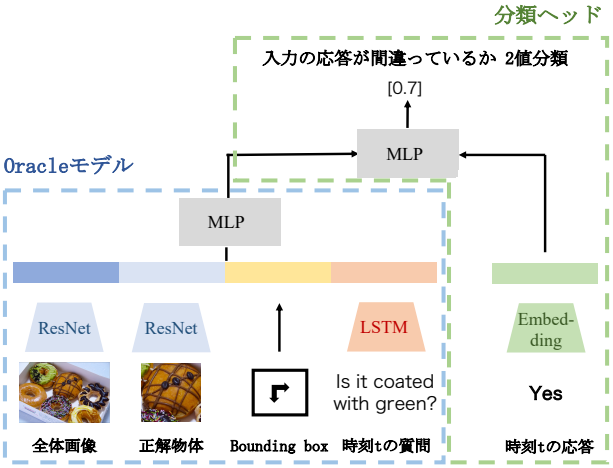


**図 4** 比較モデルの概要図

**表 2** 各データセットの利用方法

| データセット | 事前学習 | fine-tuning | test |
|---|---|---|---|
| Guess What?! | ✓ | - | - |
| 人工 | ✓ | - | - |
| 手動（seen） | - | ✓ (75%) | ✓ (25%) |
| 手動（unseen） | - | - | ✓ (100%) |

**データセット**　手動データセットが小規模である問題点に対し，「staus = Success」のラベル付けされたデータセットを用いて，新しく**人工データセット**を構築した．具体的には，人間の応答の「Yes」と「No」をランダムに反転することで，人工的に間違いのある応答を作りだし，約 11 万対話の規模のデータセットとした．

本実験では上記の人工データセットに加え，Guess What?!データセット [2]，手動データセットを含めた 3 種類のデータセットを用いた．これらデータセットの利用方法を表 2 に示す．Guess What?!，人工データセットは事前学習に，手動データセットは，fine-tuning 時と評価に用いた．手動データセットは，訓練時と同じ画像で構成される seen データセットと異なる画像で構成される unseen データセットに分割した．さらに，seen データセットに関しては fine-tuning 用と評価用に分割した．unseen データセットは全て評価に用いた．ここで，評価時に用いた seen，unseen データセットは共に，対話テキストに関しては未知のものである．

2) 対象の単語が含まれる場合に，特定の質問タイプに分類する方式．“left’ が含まれている質問は Spatial に分類される．

3) 本研究は**人間の応答の間違い指摘**，間違いの応答が**正例**で，正しい応答が**負例**となることに注意する必要がある．

本手法では，あえて対話履歴をモデルの入力として加えていない．この理由は，人工データセットに生じる不自然な対話の流れ（付録 A）をモデルが学習することを避けるためであり，本研究における提案手法の一つである人工データセットを用いる上でのモデルへの制約だといえる．

**評価指標**　間違い指摘の分類性能の評価には，F 値を用いた（再現率，適合率は付録 D を参照）．この評価指標を用いたのは，人間は応答ミスをあまりせず，テストデータセットは正例数が負例数よりも少ない不均衡データセットであるためである．

### 3.1　各実験の詳細

**実験 1**　人間が応答で間違えたデータが少ないという問題に対処するため，人工データセットによる事前学習の有効性の検証実験を行った．具体的には，1) 手動データセットのみの学習方法，2) Guess What?!データセットを用いて Oracle モデルとして学習した後に，手動データセットによる分類ヘッドの学習（転移学習）を行う学習方法，3) 人工データセットで事前学習し手動データセットを用いて fine-tuning する学習方法の 3 つを比較した．

ここで，人間は応答ミスをあまりしないため，手動データセットは正例の数が負例の数よりも少ない不均衡データセットになっている．したがって，fine-tuning 用の手動データセットは，正例と負例の数が等しくなるようにオーバーサンプリングした．

**実験 2**　2 章の分析結果が人間の応答ミスに対する指摘の正確性を向上に繋がるかを検証した．具体的には，**応答時刻**と**質問タイプ**をそれぞれ新しく入力に加えるモデル（付録 C）を構築し，各モデルの間違い指摘性能を比較する．その際 fine-tuning 時に用いるサンプルは，各質問タイプの数が同じになるようにオーバーサンプリングした．

## 4　実験結果

実験 1 の結果を表 3 に示す．人工データセットで事前学習する方法が最も良い F 値が得られ，有効的な学習方法であることがわかる．

実験 2 の結果を表 4 に示す．モデルの構造が事前学習段階から異なるため，事前学習のみの結果も記載している．次に，応答時刻を入力に追加したモデルにおける，最終時刻の応答とそれ以外の応答での比較の結果を表 5 に示す．

表 4 から，seen においては追加する入力がないモデルが一番良く，unseen においては応答時刻，質問タイプのどちらを加えても良い結果となった．また，表 5 より，unseen に対して，最終時刻以外の応答に対する F 値を下げることなく，最終時刻の応答に対する F 値を上げることができた．

これらの結果から，seen は既知の画像であるため，視覚情報を上手く捉えることができ，新しい入力がかえってモデルに不必要な複雑さを与えてしまい，F 値が下がったと考えられる．unseen では視覚情報を上手く捉えることができず，質問タイプが良いヒントになったと考えられる．また，応答時刻に関してもデータの分布をただ学習するのではなく，良いヒントとしてモデルが利用できていることが示唆される．

**表 3**　各学習方法における間違い指摘の F 値比較

| 学習 step1 | 学習 step2 | seen | unseen |
|---|---|---|---|
| なし | 手動 | 0.730 | 0.368 |
| Guess What?! | 手動 | 0.354 | 0.269 |
| 人工 | 手動 | **0.811** | **0.482** |

**表 4**　モデルの追加入力の F 値結果

| 追加入力 | 学習方法 | seen | unseen |
|---|---|---|---|
| なし | 事前学習のみ | 0.397 | 0.383 |
| | 事前学習+fine-tuning | **0.811** | 0.482 |
| 応答時刻 | 事前学習のみ | 0.175 | 0.307 |
| | 事前学習+fine-tuning | 0.718 | 0.514 |
| 質問タイプ | 事前学習のみ | 0.392 | 0.412 |
| | 事前学習+fine-tuning | 0.743 | **0.527** |

**表 5**　最終時刻の応答と最終時刻以外における応答のモデルの F 値比較

| 追加入力 | 最終時刻の応答 | | 最終時刻以外の応答 | |
|---|---|---|---|---|
| | seen | unseen | seen | unseen |
| なし | **0.875** | 0.548 | **0.714** | 0.406 |
| 応答時刻 | 0.789 | **0.608** | 0.609 | 0.406 |

## 5　おわりに

本研究では，Guess What?!における人間の応答に含まれる間違いの指摘において，人工データセットによる事前学習が有効であること，応答時刻・質問タイプの追加入力がモデルの精度向上に部分的に貢献することが示唆された．本研究は Guess What?!という限られたドメインにおける間違い指摘のみを扱ったため，今後の展望として，本研究で得られた知見を他の種類の視覚情報を用いたタスク指向型対話に対しても検証することを考えている．

## 謝辞

本研究は JSPS 科研費 (19H04137，21H05054，21K17806) の助成を受けたものです．

## 参考文献

[1] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In **CVPR**, 2017.

[2] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In **CVPR**, 2017.

[3] Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In **EMNLP**, 2021.

[4] Abhishek Das, Satwik Kottur, Stefan Lee José M. F. Moura, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In **ICCV**, 2017.

[5] Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In **ICCV**, 2017.

[6] Tao Tu, Qing Ping, Govind Thattai, Gokhan Tur, and Prem Natarajan. Learning better visual dialog agents with pretrained visual-linguistic representation. In **CVPR**, 2021.

[7] Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In **ICCV**, 2021.

[8] "Anthony Sicilia, Tristan Maidment, Pat Healy, and Malihe Alikhani". Modeling non-cooperative dialogue: Theoretical and empirical insights. **arXiv preprint arXiv:2207.07255**, 2022.

[9] Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. They are not all alike: Answering different spatial questions requires different grounding strategies. In **SpLU**, 2020.

[10] Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In **NAACL**, 2019.

## A　対話の流れが不自然になる例

1. Is it a food item? **Yes**
2. Is it a plate? No
3. Is it a glass? No
4. Is it a table? Yes

**図 5**　人工データセットにおける対話の流れが不自然になる例．1 番目の応答が反転により間違いとなっている．

図 5 は，時刻 1 で「食べ物」であることが確定しているのにも関わらず，時刻 2 で「Is it a plate?」と聞くような不自然な対話の流れになっている．実際の対話の場合，「Is it a banana?」などといった食べ物に関する質問が続くはずである．

## B　不均衡データにおける評価指標

正例（陽性）の割合が少ないテストデータにおいて，正解率（= 正解数 / サンプル数）は適切に評価できない．理由は，モデルが単純にサンプル数の多い負例を予測するように学習すれば，全く正例を予測できないにも関わらず，高い正解率を誇る良いモデルとして評価されてしまうためである．そこで，不均衡データにおける評価指標の一つとして再現率・適合率により定義される F 値がある．真陽性（陽性ラベルを正しく陽性と予測），真陰性（陰性ラベルを正しく陰性と予測），偽陽性（陰性ラベルを誤って陽性と予測），偽陰性（陽性ラベルを誤って陰性と予測）として，再現率・適合率は以下で計算される．

$$\text{再現率} = \frac{\text{真陽性数}}{\text{真陽性数} + \text{偽陰性数}} \tag{2}$$

$$\text{適合率} = \frac{\text{真陽性数}}{\text{真陽性数} + \text{偽陽性数}} \tag{3}$$

正例（陽性）の割合が少ないテストデータにおいて，モデルが単純に「正例」と多く予測するように学習すれば再現率は高く適合率は低くなり，反対に「負例」と多く予測するように学習すれば適合率は高くなり再現率は低くなる．このようなモデルを適切に低く評価するために，再現率と適合値の調和平均を取った F 値がある．

$$\text{F 値} = 2 \times \frac{\text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{4}$$

## C　実験 2 における追加入力モデル

実験 2 で用いた，**質問タイプ**と**応答時刻**を新しく入力として加えたモデルの概要図を，図 6，7 にそれぞれ示す．

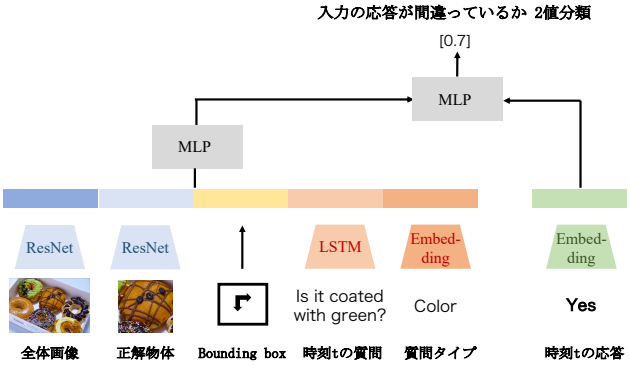


**図 6**　質問タイプを入力に新しく加えたモデル

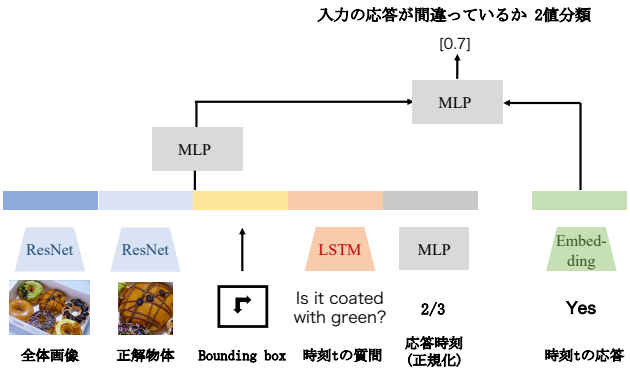


**図 7**　応答時刻を入力に新しく加えたモデル

## D　実験結果の補足

実験 1，2 の再現率・適合率の結果をそれぞれ表 6, 7 に示す．

**表 6**　各学習方法における，間違い指摘の精度比較

| 学習 step1 | 学習 step2 | seen | | unseen | |
|---|---|---|---|---|---|
| | | 再現率 | 適合率 | 再現率 | 適合率 |
| なし | 手動 | **0.92** | 0.605 | 0.459 | 0.308 |
| Guess What?! | 手動 | 0.520 | 0.268 | 0.377 | 0.209 |
| 人工 | 手動 | 0.860 | **0.768** | **0.541** | **0.434** |

**表 7**　モデルの追加入力の F 値結果

| 追加入力 | 学習方法 | seen | | unseen | |
|---|---|---|---|---|---|
| | | 再現率 | 適合率 | 再現率 | 適合率 |
| なし | 事前学習のみ | 0.580 | 0.302 | 0.508 | 0.307 |
| | +fine-tuning | **0.860** | **0.768** | 0.541 | 0.434 |
| 応答時刻 | 事前学習のみ | 0.220 | 0.145 | 0.344 | 0.276 |
| | +fine-tuning | 0.840 | 0.627 | 0.623 | 0.437 |
| 質問タイプ | 事前学習のみ | 0.580 | 0.296 | 0.574 | 0.321 |
| | +fine-tuning | 0.840 | 0.667 | **0.639** | **0.448** |