# クラス間で良く分離された説明文を自動生成する大規模視覚言語モデルを用いた対照学習による歩行者年齢層認識

尾崎匠 (鳥取大)　栗林英範 (グローリー)　井上路子 (鳥取大)　西山正志 (鳥取大)

鳥取大学 Tottori University

# 研究背景

**防犯カメラに映る人物の全身画像**を用いた**年齢層クラス**を認識する技術が期待されている

特定の範囲に分けられた年齢が属する集合

例
- 15歳以下
- 16歳以上 30歳以下
- 31歳以上 45歳以下
- 46歳以上 60歳以下
- 61歳以上

**各年齢層クラスの中で人物の向きや解像度などの要因により多様に変動**

顔が鮮明に写っている場合に適用できる顔画像を用いた年齢層認識の手法は
全身画像において適用できるとは言えない

見え方が多様に変動する全身画像から年齢層クラスを高精度に認識できる手法が求められる

# 年齢層認識の既存手法の概要

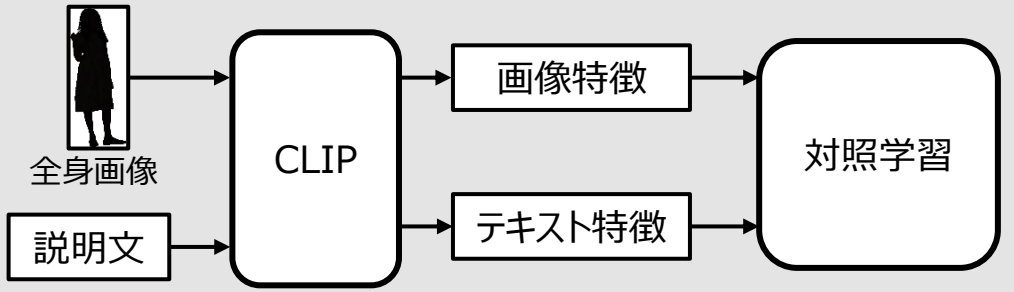
近年では**画像特徴**と**テキスト特徴**とを同時に扱う認識手法が登場

## 既存手法1 (PromptPAR)

[X.Wang et al., TCSVT(2024)]

- 画像とテキストとを用いるCLIPモデル※1が基盤
- 各属性クラスの説明文からテキスト特徴を抽出し対照学習

※1[R.Alec et al., ICML (2021)]



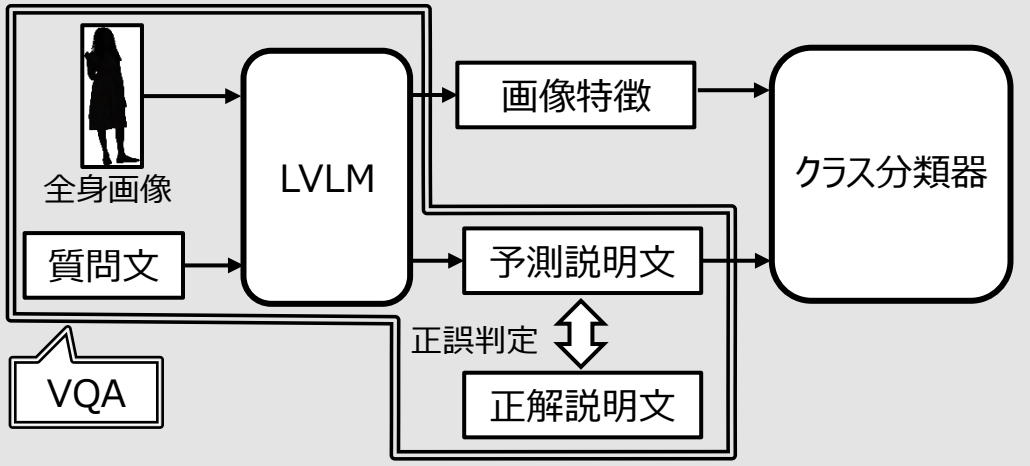

## 既存手法2 (LLMPAR)

[J.Jin et al., arXiv:2408.09720 (2024)]

- VQAを行う大規模視覚言語モデル (MiniGPT-4※2)が基盤
- VQAによって生成された文章を認識に使用し各属性クラスの説明文を正解としてモデルを訓練

※2[C.Jun et al., arXiv:2310.09478 (2023)]

LVLM (Large Vision Language Model)：大規模視覚言語モデル




画像特徴とテキスト特徴とを用いる手法は画像特徴のみを用いる手法※3と比較して精度向上

※3[F.Xinwen et al., TCSVT(2023)][J.Jian et al., ICCV(2021)] [W. Junyi et al., Pattern Recognition (2022)] [C. Lin et al., Neural Computing and Applications(2022)]

# 既存手法のPromptPARやLLMPARの課題

## 課題1：年齢層クラス間で良く分離された説明が設計されていない

PromptPAR | LLMPAR

- 年齢層クラスを数単語のみで表した説明文を使用
  (例：A pedestrian whose age is 16 or younger.)
- 説明文が年齢層クラス間で良く分離されていなければテキスト特徴も良く分離されず認識精度が低下

## 課題2：説明文を手動で作成

PromptPAR | LLMPAR

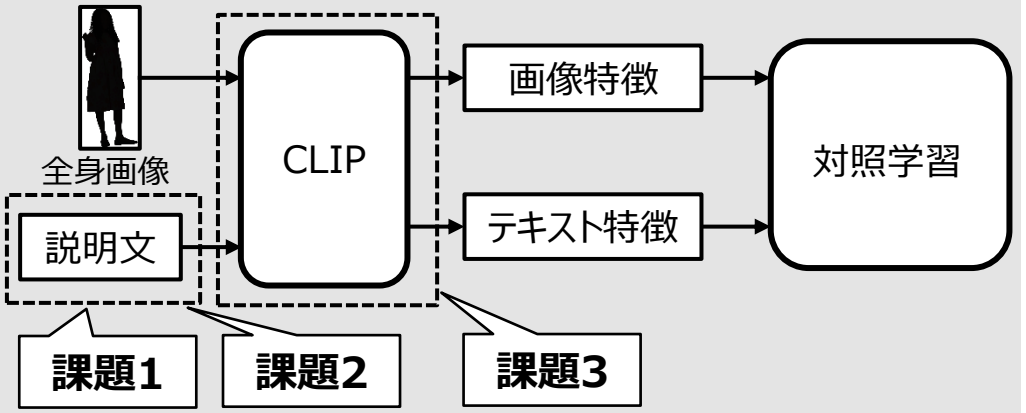
- 手動で説明文を作成し認識精度を確認するため試行錯誤が必要
- 説明文の単語数を増やすと試行錯誤の手間が多くかかる

## 課題3：最新の大規模視覚言語モデルから抽出された特徴を対照学習に用いていない

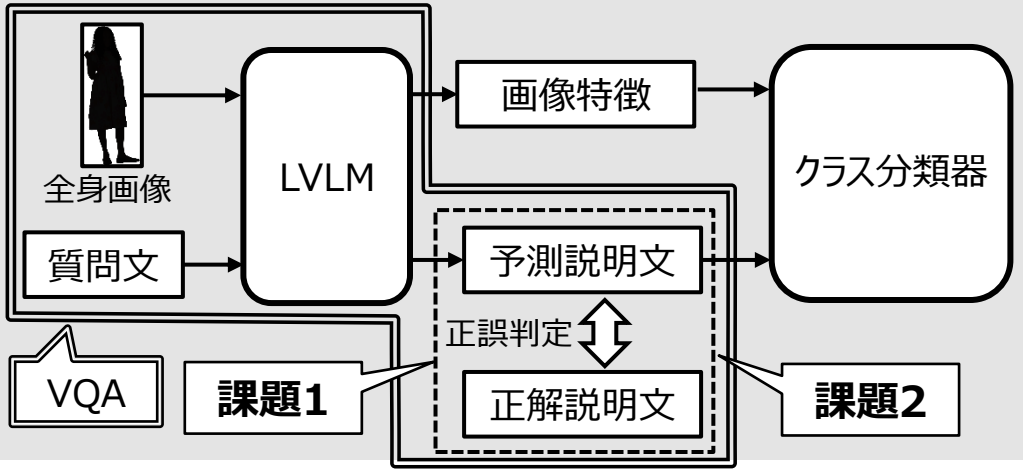
PromptPAR

- 特徴抽出にCLIPモデル※を用いる ※[R.Alec et al., ICML (2021)]
- CLIPモデルは入力できるトークン数に限りがある
  ➡ 単語数の多い説明文を入力できない場合がある

PromptPAR [X.Wang et al., TCSVT(2024)]

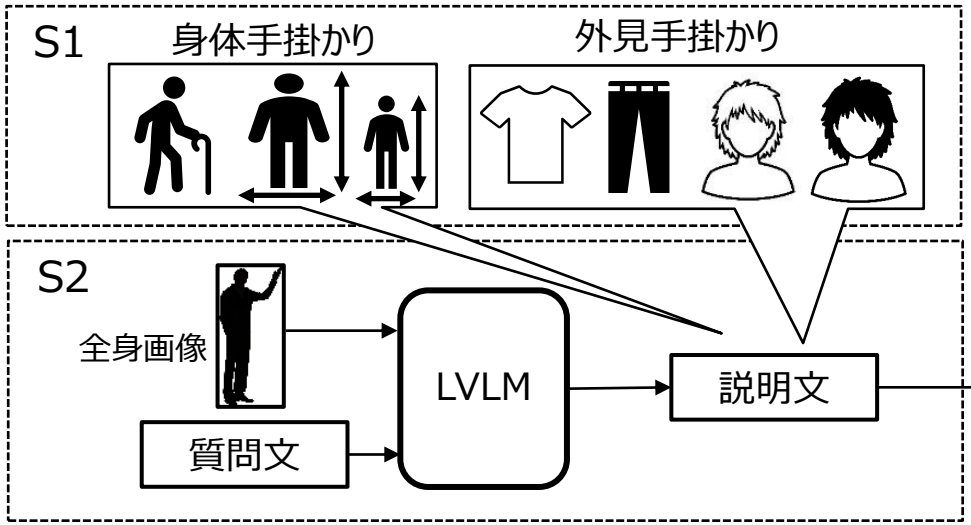



LLMPAR [J.Jin et al., arXiv:2408.09720 (2024)]

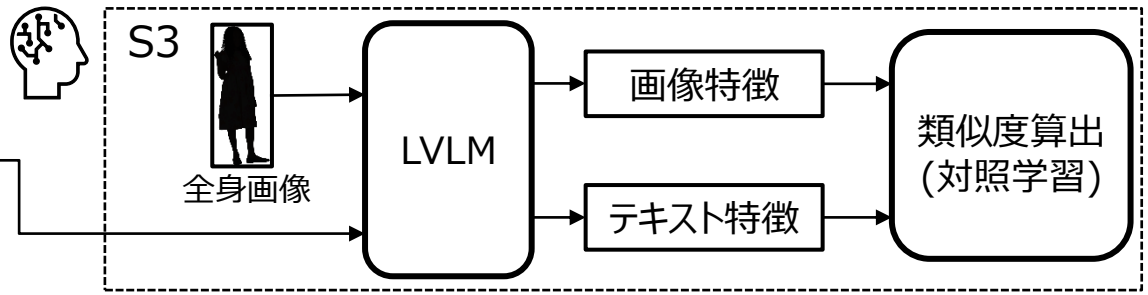


# 研究目的

**歩行者の全身画像から年齢層クラスを精度よく手間なく認識するため**
**年齢層クラス間で良く分離された説明文を設計し**
**説明文を自動生成しその説明文を用いて対照学習を行う**




貢献1：年齢層クラスを表す説明文を**クラス間で良く分離**されるように設計




貢献2：大規模視覚言語モデル用いることで説明文を**試行錯誤の手間なく**自動生成

貢献3：大規模視覚言語モデルを用いてクラス間で**良く分離されたテキスト特徴を抽出し**
**認識精度向上**のため画像特徴を合わせる対照学習

(RAPv2.0, PETA, MSP60k)
公開データセットにおいて既存手法と比較して認識精度が向上

# 提案手法の全体の流れ



S1.年齢層クラス間で良く分離された説明文の設計

身体手掛かり (身長，体型など)

外見手掛かり (服装，髪色など)

年齢層クラス$C_1$の説明文$P_1$

Their clothing styles vary but generally reflect practical and comfortable fashion choices suitable for children. This includes items such as t-shirts, shorts, sneakers, and jackets, which allow for ease of movement during play and daily activities. Additionally, children are in the early stages of growth and tend to have smaller physiques, reflecting their developmental phase.

年齢層クラス$C_N$の説明文$P_N$

The images depict people who appear to be elderly, with common characteristics such as bent hips and gray hair. They are dressed casually, appropriate for everyday activities such as shopping and running errands.
Clothing ranges from jackets and coats to shirts and pants, indicating a variety of indoor or outdoor activities.

S2. 説明文の自動生成 (大規模視覚言語モデルを用いたVQA)

年齢層クラス$C_1$の質問文$Q_1$

年齢層クラス$C_N$の質問文$Q_N$

年齢層クラス$C_1$の全身画像群$I_1$ ($M$枚)

年齢層クラス$C_N$の全身画像群$I_N$ ($M$枚)

VQAによる説明文の自動生成

LVLM

The age of the pedestrian in these images is [$C_n$]. Describe common characteristics in detail that can be observed in these images.

S3. 画像特徴とテキスト特徴とで対照学習 (大規模視覚言語モデルを用いた特徴抽出)

内積

Frozen

Tuning

特徴抽出

LVLM

Language Model

Image Encoder

全身画像$I$

$f_1^l$

$f_N^l$

各年齢層クラスのテキスト特徴

$f^v$

画像特徴

特徴間で類似度算出

Projector

Adapter

$\tilde{f}^v \cdot \tilde{f}_n^l$

Loss (Cross Entropy)

年齢層クラス

尤度

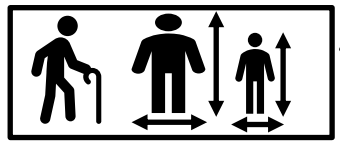
$C_1$ 0.2

$C_N$ 0.1

# 提案手法の全体の流れ



S1.年齢層クラス間で良く分離された説明文の設計



身体手掛かり
(身長，体型など)

外見手掛かり
(服装，髪色など)

年齢層クラス$C_1$の説明文$P_1$

Their clothing styles vary but generally reflect practical and comfortable fashion choices suitable for children. This includes items such as t-shirts, shorts, sneakers, and jackets, which allow for ease of movement during play and daily activities. Additionally, children are in the early stages of growth and tend to have smaller physiques, reflecting their developmental phase.

⋮

年齢層クラス$C_N$の説明文$P_N$

The images depict people who appear to be elderly, with common characteristics such as bent hips and gray hair. They are dressed casually, appropriate for everyday activities such as shopping and running errands.
Clothing ranges from jackets and coats to shirts and pants, indicating a variety of indoor or outdoor activities.

S2. 説明文の自動生成
(大規模視覚言語モデルを用いたVQA)

VQAによる説明文の自動生成

年齢層クラス$C_1$の質問文$Q_1$

年齢層クラス$C_1$の全身画像群$I_1$ ($M$枚)

⋮

年齢層クラス$C_N$の質問文$Q_N$

年齢層クラス$C_N$の全身画像群$I_N$ ($M$枚)

The age of the pedestrian in these images is [$C_n$]. Describe common characteristics in detail that can be observed in these images.

LVLM

**S1について説明**

S3. 画像特徴とテキスト特徴とで対照学習
(大規模視覚言語モデルを用いた特徴抽出)

内積

Frozen

Tuning

特徴間で対照学習

Language Model

$f_1^l$

⋮

$f_N^l$

各年齢層クラスのテキスト特徴

Projector

Loss
(Cross Entropy)

全身画像$I$

Image Encoder

$f^v$

画像特徴

Adapter

$\tilde{f}^v \cdot \tilde{f}_n^l$

年齢層クラス

尤度

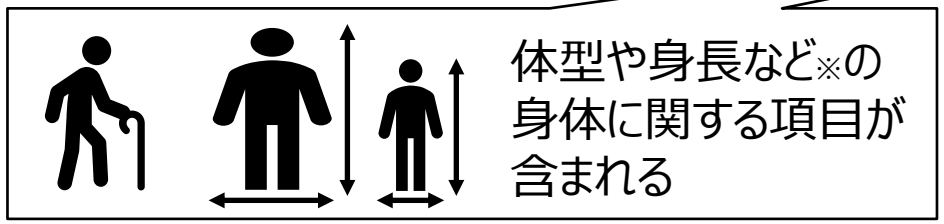
$C_1$ 0.2

⋮

$C_N$ 0.1

# S1:年齢層クラス間で良く分離された説明文の設計

- テキスト特徴を良く分離するためには説明文が良く分離されている必要がある
- **説明文の設計方針**：各年齢層クラスを端的に表す**身体手掛かり**や**外見手掛かり**について単語数を増大させて詳細に記述



体型や身長など※の身体に関する項目が含まれる

具体例：腰が曲がっている，太っている，身長が高い

服装や髪色など※の外見に関する項目が含まれる

具体例：Tシャツ，ジーンズ，白髪，黒髪

※[D.Antitza et al., Multimedia Tools and Applications (2011)]

- **提案手法の考え方**：各年齢層クラスを端的に表す項目を説明文に含めることが好ましい
  - なお，各年齢層クラスにおいてすべての項目を説明文に含める必要はない
- **提案手法における説明文の例**
  - 15歳以下の場合：低い身長で，小柄な体型で，カジュアルな服装である.
  - 61歳以上の場合：曲がった腰で，白髪で，薄毛である.

年齢層クラス間で説明文に大きな差が生じ説明文がクラス間で良く分離される

# 提案手法の全体の流れ S2について説明



S1.年齢層クラス間で良く分離された説明文の設計

身体手掛かり (身長，体型など)

外見手掛かり (服装，髪色など)

年齢層クラス$C_1$の説明文$P_1$

Their clothing styles vary but generally reflect practical and comfortable fashion choices suitable for children. This includes items such as t-shirts, shorts, sneakers, and jackets, w… ties. Additionally, children are in the early stages of growth and tend to have smaller phy…

年齢層クラス$C_N$の説明文$P_N$

The images depict peop… ch as bent hips and gray hair. They are dressed casually, appropriate for everyday activities such as shopping and running errands.
Clothing ranges from jackets and coats to shirts and pants, indicating a variety of indoor or outdoor activities.

**S2について説明**

S2. 説明文の自動生成
(大規模視覚言語モデルを用いたVQA)

VQAによる説明文の自動生成

年齢層クラス$C_1$の質問文$Q_1$

年齢層クラス$C_1$の全身画像群$I_1$ ($M$枚)

年齢層クラス$C_N$の質問文$Q_N$

年齢層クラス$C_N$の全身画像群$I_N$ ($M$枚)

LVLM

The age of the pedestrian in these images is [$C_n$]. Describe common characteristics in detail that can be observed in these images.

S3. 画像特徴とテキスト特徴とで対照学習
(大規模視覚言語モデルを用いた特徴抽出)

内積

Frozen

Tuning

特徴抽出

特徴間で対照学習

LVLM

Language Model

Image Encoder

$f_1^l$

$f_N^l$

各年齢層クラスのテキスト特徴

$f^v$

画像特徴

全身画像$I$

Projector

Adapter

$\tilde{f}^v \cdot \tilde{f}_n^l$

Loss
(Cross Entropy)

年齢層クラス

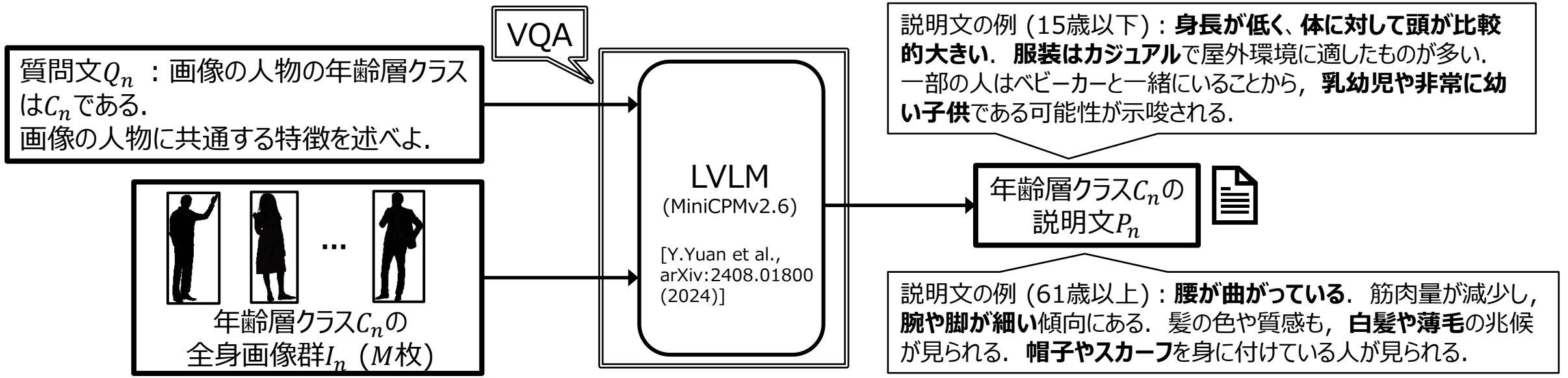
尤度

$C_1$ 0.2

$C_N$ 0.1

# S2: 説明文の自動生成

- 詳細な説明文をVQAに回答させるには質問文が重要
- **単純な考え方**：身体手掛かりや外見手掛かりの項目を列挙した質問文
  - ✗ 項目をすべて列挙するには手間がかかる
  - ✗ 問い合わせたすべての項目に対しVQAが回答するとは限らない（例：向きや解像度で読み取れない項目）
- **提案手法の考え方**：手掛かりについて**項目を陽に指定しない簡潔な質問文**
  - VQA内の大規模視覚言語モデルに内在する年齢層クラスに関する項目を自然に表出




詳細な説明文が自動生成され試行錯誤の手間がなくなる

# 提案手法の全体の流れ S3について説明



S1.年齢層クラス間で良く分離された説明文の設計

身体手掛かり (身長，体型など)

外見手掛かり (服装，髪色など)

年齢層クラス$C_1$の説明文$P_1$

Their clothing styles vary but generally reflect practical and comfortable fashion choices suitable for children. This includes items such as t-shirts, shorts, sneakers, and jackets, w… ties. Additionally, children are in the early stages of growth and tend to have smaller phy…

年齢層クラス$C_N$の説明文$P_N$

The images depict peop… ch as bent hips and gray hair. They are dressed casually, appropriate for everyday activities such as shopping and running errands.
Clothing ranges from jackets and coats to shirts and pants, indicating a variety of indoor or outdoor activities.

**S3について説明**

S2. 説明文の自動生成 (大規模視覚言語モデルを用いたVQA)

VQAによる説明文の自動生成

年齢層クラス$C_1$の質問文$Q_1$

年齢層クラス$C_1$の全身画像群$I_1$ ($M$枚)

年齢層クラス$C_N$の質問文$Q_N$

年齢層クラス$C_N$の全身画像群$I_N$ ($M$枚)

LVLM

The age of the pedestrian in these images is [$C_n$]. Describe common characteristics in detail that can be observed in these images.

S3. 画像特徴とテキスト特徴とで対照学習 (大規模視覚言語モデルを用いた特徴抽出)

内積

Frozen

Tuning

特徴抽出

LVLM

Language Model

Image Encoder

全身画像$I$

$f_1^l$

$f_N^l$

各年齢層クラスのテキスト特徴

$f^v$

画像特徴

特徴間で対照学習

Projector

Adapter

$\tilde{f}^v \cdot \tilde{f}_n^l$

Loss (Cross Entropy)

年齢層クラス

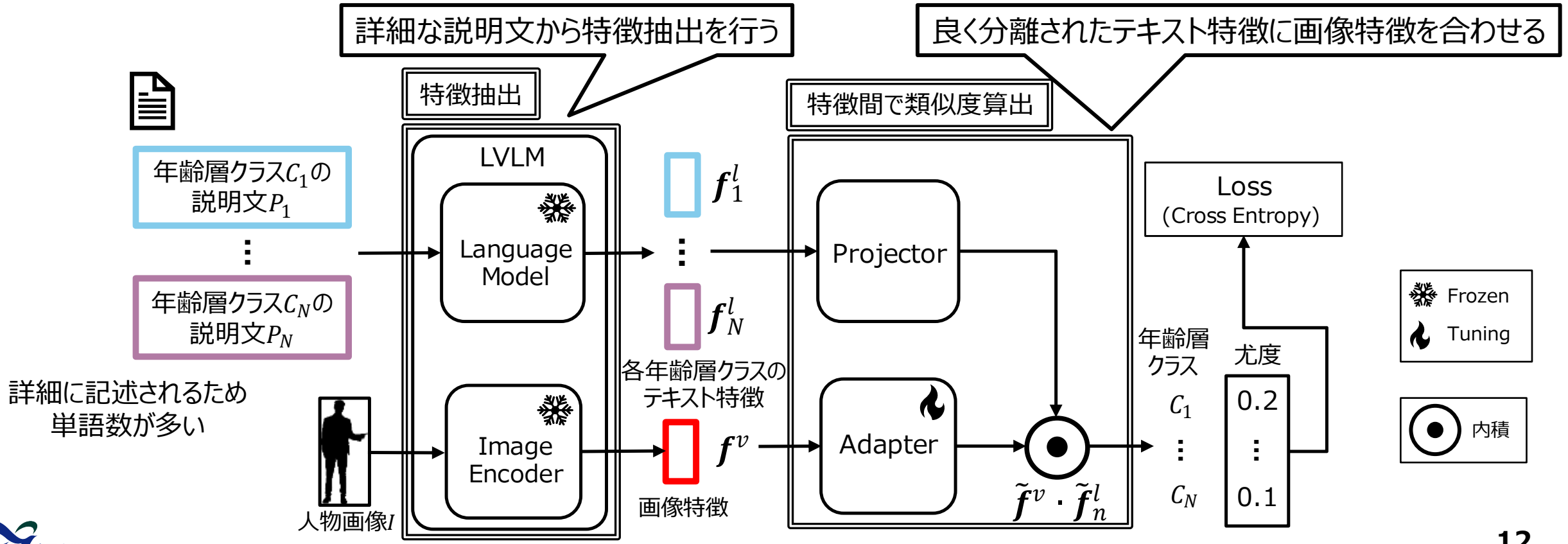
尤度

$C_1$ 0.2

$C_N$ 0.1

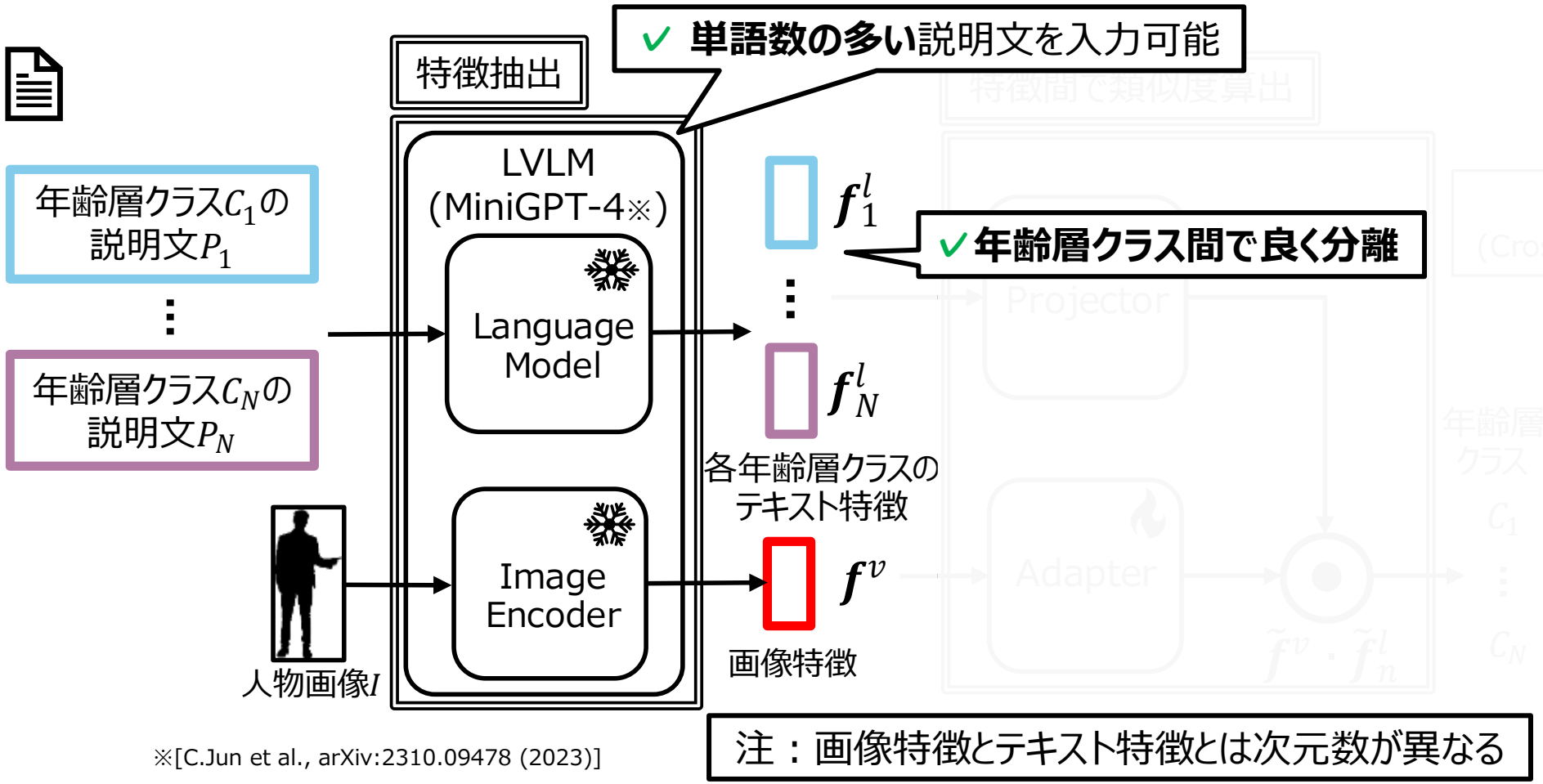
# S3: 画像特徴とテキスト特徴とで対照学習

- 詳細な説明文から抽出されたテキスト特徴に画像特徴を合わせる
- **単純な考え方**：CLIPモデル※をファインチューニング ※[R.Alec et al., ICML (2021)]
  - ✗ 単語数の多い詳細な説明文を入力できず，年齢層クラス間で良く分離されたテキスト特徴を抽出できない
- **提案手法の考え方**：大規模視覚言語モデルから抽出された特徴で対照学習を行う



# S3: 画像特徴とテキスト特徴とで対照学習（特徴抽出）

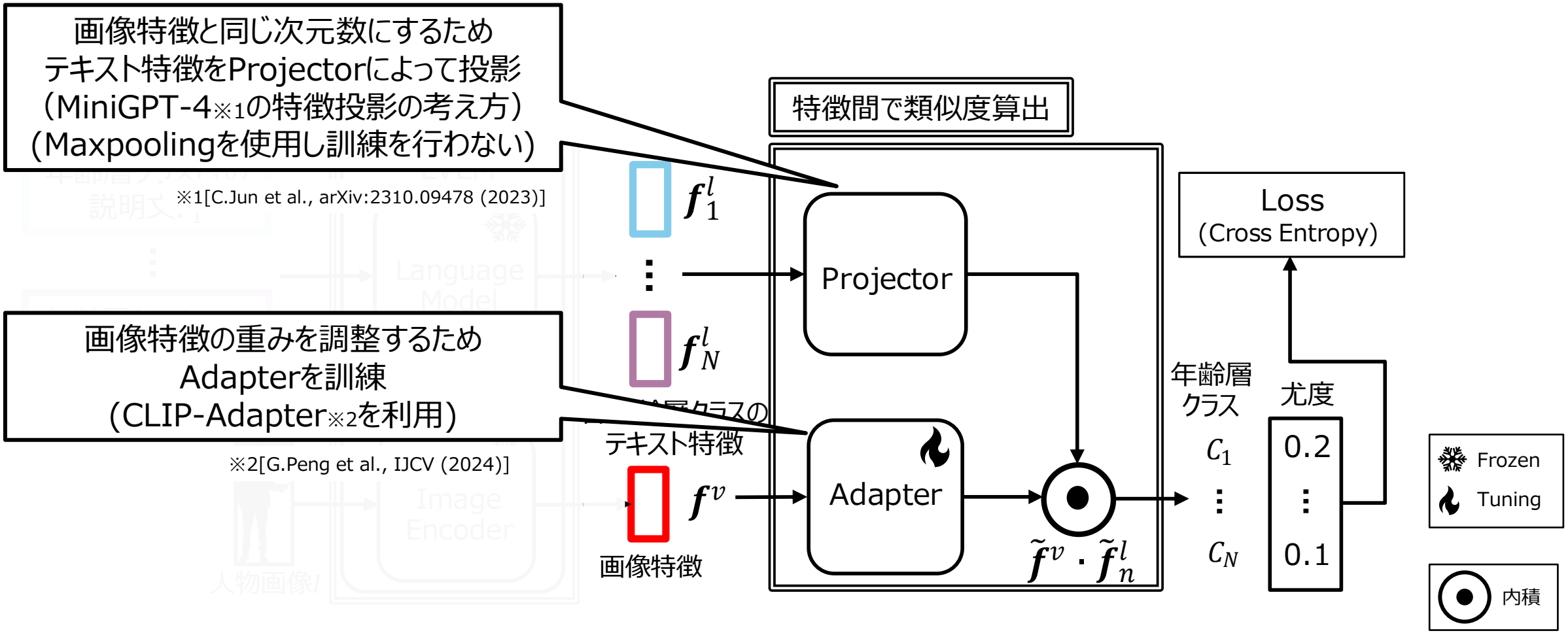
- 詳細な説明文から特徴抽出するため入力トークン数が多い大規模視覚言語モデルを適用




※[C.Jun et al., arXiv:2310.09478 (2023)]

# S3: 画像特徴とテキスト特徴とで対照学習（類似度算出）

- テキスト特徴に画像特徴を合わせるためProjectorで次元調整しAdapterで重み調整




詳細な説明文から抽出した年齢層クラス間で良く分離されたテキスト特徴に
画像特徴を合わせる対照学習を行うことで認識精度を向上

# 評価に使用するデータセット

公開データセット

- RAPv2.0 [L.Dangwei et al., IEEE TIP (2018)]
- PETA [D.Yubin et al., ACM (2014)]
- MSP60k [J.Jin et al., arXiv:2408.09720 (2024)]

画像枚数の調整

- 年齢層クラス間で画像枚数の偏りがある
- 画像枚数が少ないクラスをアップサンプリング
  - どの年齢層も同じ確率で出現することを想定
- 評価に使用する画像枚数：47,215枚

実験条件

- 評価方法：5-Fold cross validation
  (データセットを5分割し，4つを訓練1つを予測)
- 説明文を自動生成するモデル：MiniCPM-v2.6
  [Y.Yuan et al., arXiv:2408.01800 (2024)]
- 画像特徴とテキスト特徴とを抽出するモデル：MiniGPT-4
  [C.Jun et al., arXiv:2310.09478 (2023)]

全身画像の例

RAPv2.0


PETA


MSP60k

# 手間と認識精度の比較

| 手法 | 特徴 | 説明文 | | 対照学習 | 試行錯誤の手間 | 年齢層クラスの認識精度 (%) |
|---|---|---|---|---|---|---|
| | | 内容 | 自動生成 | | | |
| PARFormer [F. Xinwen et al., TCSVT (2023)] | 画像のみ | – | – | – | – | 70.4 |
| SSC [J.Jia et al., ICCV (2021)] | 画像のみ | – | – | – | – | 65.3 |
| CLIP (Zero-Shot) [R.Alec et al., ICML (2021)] | 画像とテキスト | 簡略 | | ✓ | あり | 35.5 |
| CLIP-Adapter [G.Peng et al., IJCV (2024)] | 画像とテキスト | 簡略 | | ✓ | あり | 70.0 |
| CoOp [Z.Kaiyang et al., IJCV (2022)] | 画像とテキスト | 簡略 | | ✓ | あり | 67.7 |
| Visual-Prompt [H.Bahng et al.,arXiv:2203.17274 (2022)] | 画像とテキスト | 簡略 | | ✓ | あり | 54.9 |
| VQA [C.Santana et al., CAIP (2023)] | 画像とテキスト | 簡略 | | | あり | 44.4 |
| PromptPAR [X.Wang et al., TCSVT(2024)] | 画像とテキスト | 簡略 | | ✓ | あり | 74.6 |
| LLMPAR [J.Jin et al., arXiv:2408.09720 (2024)] | 画像とテキスト | 簡略 | | | あり | 51.9 |
| **提案手法** | 画像とテキスト | 詳細 | ✓ | ✓ | **なし** | **79.9** |

**提案手法は試行錯誤の手間を削減し認識精度を改善**

–：対象外

# テキスト特徴の分離度合いの比較

- 身体手掛かりまたは外見手掛かりについて単語数を増大させて詳細に記述した説明文が年齢層クラス間で良く分離されているかを確認
- 説明文から抽出したテキスト特徴の年齢層クラス間の**分離度合い**を算出
  - **分離度合い**：各年齢層クラスに1つずつ存在するテキスト特徴の分散を算出

(15歳以下)：手足や頭の比率など，成人の体格にはまだ達していない成長過程にある人物の身長や体格，子供によく見られるカジュアルな服装（Tシャツやショートパンツ，ズボンなど)が見られる.

A pedestrian whose age is [Label].

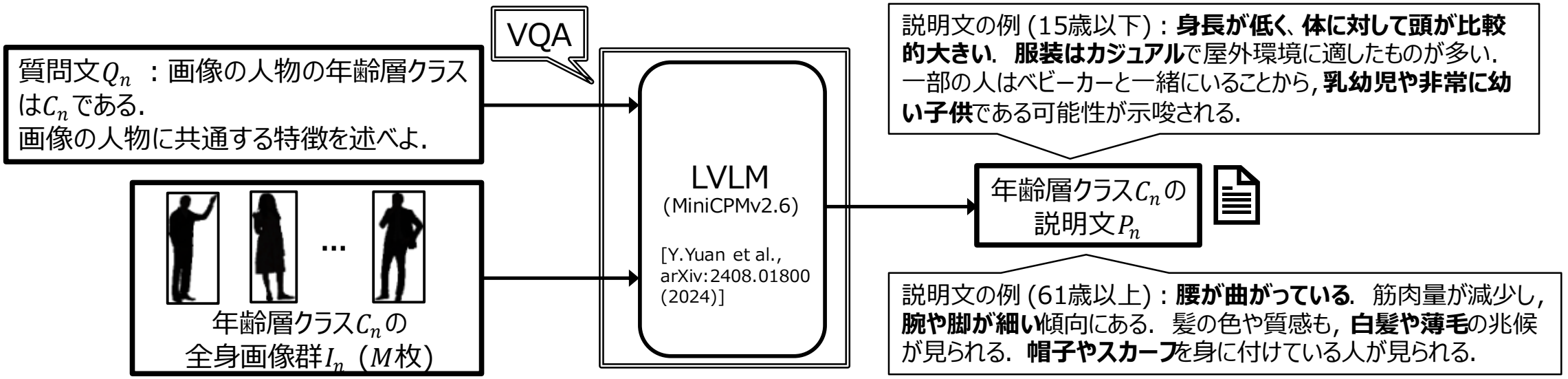
A [Label] person

| 説明文 | 分離度合い↑ |
|---|---|
| **提案手法** | **0.72** |
| PromptPAR [X.Wang et al., TCSVT(2024)] | 0.14 |
| LLMPAR [J.Jin et al., arXiv:2408.09720 (2024)] | 0.29 |

**身体手掛かりまたは外見手掛かりについて詳細に記述することで**
**テキスト特徴の分離度合いが年齢層クラス間で向上**

# 説明文生成に使用する画像枚数の比較

- 説明文を生成する際に何枚の画像を使用すればテキスト特徴が良く分離されるかを確認




全身画像の枚数を比較

| 画像枚数 | 認識精度 (%) |
|---|---|
| 1 | 78.3 |
| 10 | 79.6 |
| 100 | 79.9 |

**画像枚数を10枚以上に増やすと認識精度が安定**

# まとめ

**歩行者の全身画像から年齢層クラスを精度よく手間なく認識するため**
**年齢層クラス間で良く分離された説明文を設計し**
**説明文を自動生成しその説明文を用いて対照学習を行う手法を提案した**

- S1. 年齢層クラスを表す説明文を**クラス間で良く分離**されるように設計
- S2. 大規模視覚言語モデル用いることで説明文を**試行錯誤の手間なく**自動生成
- S3. 大規模視覚言語モデルを用いて**クラス間で良く分離されたテキスト特徴を抽出し認識精度向上**のため画像特徴を合わせる対照学習

## 今後の課題

- 説明文に含める項目の検討
- 認識精度が保障される質問文設計の方法論の構築
- 計算コスト削減