

# 第1章 提案手法

本章では、知識ネットワークを用いてシングラフを生成する提案手法について述べる。提案手法の特徴は、LLM が学習した常識的知識に基づいて**物体間関係のネットワークとして構築し**、入力画像に対して**物体認識によって得られる観測情報**を手がかりに**ネットワーク探索を行う**点にある。

具体的には、まず物体ペアに対する関係語・位置関係・接触関係を LLM により大量生成し、それらの出現頻度を用いて関係グラフを構築する。次に、入力画像から物体候補の列挙と座標推定を行い、得られた物体と位置情報を探索の起点として知識ネットワークを探索することで、物理的・空間的に整合する三つ組〈主語、関係、目的語〉を推定する。最後に、推定された三つ組を統合し、画像に対応するシーングラフを構成する。

提案手法の処理フローを図??に示す。

## 1.1 知識ネットワーク構築

本節では、物体間関係に関する知識をネットワークとして構築する方法を述べる。知識ネットワークは後段のシングラフ生成における探索空間として用いられ、画像から得られる観測情報（物体名・位置関係・接触情報）に整合する関係推論を可能にする。

ネットワーク構築は大きく二段階から成る。第一に、物体ペアに対して LLM を用いて関係語を生成し、関係語に加えて位置関係（vertical/horizontal/depth）および接触関係（attach）を含む関係データを得る（第 1.1.1 項）。第二に、得られた関係データを集計し、物体・位置付き物体・関係表現・三つ組を結ぶ多層構造の有向グ

ラフとして知識ネットワークを構築する（第 1.1.2 項）.

### 1.1.1 関係語生成

本節では、物体語彙集合  $\mathcal{V}$  に対して、全ての物体ペア  $(o_i, o_j) \in \mathcal{V} \times \mathcal{V}$  を対象とし、それぞれの関係を大規模言語モデル（GPT-4o）によって生成する。生成結果は、関係語（relation）に加え、位置関係（position）および接触関係（attach）を含む JSON 形式で出力される。

本手法では、関係生成処理を多数の物体ペアに対して自動的かつ反復的に実行し、生成結果を後段のネットワーク構築に利用することを前提としている。そのため、API が公開されており利便性が高い GPT-4o を採択する。

関係語の生成には、LLM に明確な出力仕様を与えるため、指示文と出力形式を明示した以下のプロンプトを使用する。このプロンプトは、抽象的な意味関係ではなく、日常的・物理的に成立する空間関係を想定して設計している。また、プロンプト内に具体的な出力例を含めることで、LLM の出力を安定化させ、より一貫した関係文の生成を促している。なお、実際に使用したプロンプト全文は付録に示す。

### プロンプトの概要 —

現実世界の日常的な空間関係や物体の配置に焦点を当て、抽象的で孤立した記述ではなく、現実で物体的にもっともらしい関係の生成を指示

#### #指示

1. n 個の文章を生成すること
2. 各項目は以下のキーを持つ JSON 形式で出力すること
  - ”sentence”：現実的な空間関係を明確に表す文脈に基づいた文章
  - ”object”：比較対象のオブジェクトのリスト
  - ”relation”：表現される空間的関係
  - ”position”：3 次元空間でどの位置にあるか
  - ”attach”：物体が物理的に接触しているか
3. 出力全体を JSON 配列で包む
4. コメントや説明を含めない

このプロンプト設計により、モデル出力をそのまま機械処理に利用できるようになっている。以下に、生成方法の詳細を解説する。

## 2 物体を含む文の生成

まず、列挙された物体集合  $\mathcal{O}$  の全ての組み合わせ  $(o_i, o_j)$  に対して、LLM の持つ一般的な知識を活用し、「現実世界で起こりうる」関係を文形式で大量に生成する。

ここで、文の主語は必ずしも固定されず、どちらの物体を主語としてよいものとする。これにより、「机の上にりんごがある」「りんごの下に机がある」など、同一の関係を異なる視点から柔軟に表現できる。

また、単純に 2 物体を含む文を生成させるだけでは、「There is an apple on the table.」のような幾何的な記述が中心となり、日常的なシーンを反映した文が得られにくい。そこで、LLM に対し**文脈的に自然な状況を反映するヒント (contextual hints)** を含めるよう指示することで、「The book is on the desk in an office.」や

「The mug is placed next to the sink in the kitchen.」のように、現実的なシーンとして成立する文の生成を誘導する。この設計により、得られる関係データは単なる座標的関係にとどまらず、意味的に妥当な日常シーンの中で成立する関係表現も獲得される。

生成された文は”sentence”フィールドに格納し、対応する物体ペアは”object”にリスト形式で記録する。さらに、この情報をもとに”relation”に関係語を格納する。

## 位置関係の生成

関係語の生成をする際、単に2物体を含む文章を生成させるだけでは、後段の探索過程で得られる関係推論が曖昧になりやすい。例えば、「apple」と「table」の2つの物体が水平方向に位置し、正しい関係は「next to」である場合でも、「on」という関係が一般的な文脈で頻出するため、物体名のみを入力として探索を行うと「on」が誤って選択されてしまう可能性がある。つまり、それらがどのような位置情報にあるかを明示されなければ、グラフ探索時に両者を結ぶ経路を適当に特定することができず、結果として推論が偶発的で不安定なものとなってしまう。

そこで本研究では、後段（第1.2.2項）の探索において画像から得られる位置情報を有効に利用できるよう、LLMに対して文章生成と同時に2物体間の空間的な位置関係を明記させるように設計する。これにより、「apple」と「table」のような語彙的関係にとどまらず、空間的に整合する構造的情報を持つようになる。

具体的には、各文において Object1 が Object2 に対して、垂直方向 (vertical)、水平方向 (horizontal)、奥行方向 (depth) の3軸における相対的な位置を、up, down, left, right, front, back, None のいずれかで符号化する。

$$\text{vertical} \in \{\text{up}, \text{down}, \text{None}\}, \text{horizontal} \in \{\text{left}, \text{right}, \text{None}\}, \text{depth} \in \{\text{front}, \text{back}, \text{None}\}$$

例えば「The apple is on the table.」の場合、「apple」は「table」の上方向に存在するため、

`position = {vertical : down, horizontal : None, depth : None}`

として出力される。

このようにして、関係語に加え位置情報を保持することで、意味的関係と空間的配置の両面を反映したネットワーク構造を構築できる。後続のグラフ探索において、単語の出現頻度に依存しない空間的に妥当な重み付けに基づく関係推論が可能になる。

### 接触関係の生成

さらに、物体間の関係をより正確に表現するため、物理的な接触の有無を明示的に生成させる。位置情報のみでは関係性を十分に区別できない場合が多く、例えば「apple」が「table」の上に位置している場合、その関係は「on」である可能性もあれば「above」である場合もある。しかし、位置情報だけでは両者の違いを識別できず、いずれも垂直方向（vertical）が”up”であるという同一の特徴として扱われてしまう。その結果、本来「above」である関係が「on」として出力されるなど、関係語が誤って推定される可能性がある。

この問題を解消するため、本研究ではLLMに対して、生成された文が想定する状況において2物体が接触しているか否かを同時に出力させる。接触情報は”attach”に記録され、その値は”True”または”False”のいずれかで表される。基準は以下の通りである。

#### 接触判定

- ”True”：物体が実際に接している  
(例：placed on, touching, leaning, stacked, held など)
- ”False”：空間的に近くても直接の接触はない  
(例：in front of, next to, above など)

このように接触情報を明示的に付与することで、「垂直方向に上方にある」という単純な位置情報だけでなく、物理的な相互関係を含む関係性を捉えることが可能となる。得られた”attach”フィールドは、後段（第1.2.2項）のTripletRankerにおいて

て探索区間の制約条件として用いられ、物理的に矛盾しない三つ組〈主語、関係、目的語〉の推論を助ける役割を果たす。

## エラー処理

LLMによる出力は確率分布に基づく生成過程を通じて行われるため、同一のプロンプトを与えても常に完全に同一の結果が得られるとは限らない。出力形式や構造を詳細に指定した場合においても、トークン生成時の確率的揺らぎや文脈解釈の不確実性により、意図しない形式の文章や構造上の誤り、あるいは値の欠落を生じることがある。これは、LLMが厳密な規則に従って動作するルールベースモデルではなく、膨大な事例から統計的に言語パターンを学習した確率的生成モデルであることに起因する。そこで本研究では、このような生成結果の不確実性に対処するため、出力が事前に定義したフォーマットに適合しているかを自動的に検証する機構を導入する。具体的には、生成結果において必要な値が欠落している場合や、構造がフォーマット仕様と一致しない場合を検出し、それらが確認された際には再生成を自動的に実行する。これにより、生成結果の構造的信頼性を確保しつつ、プロンプトの自由度と創発性を維持したまま、安定した関係データの取得を実現している。

以下に、本研究で定義したフォーマット検証を通過し、最終的に正しい出力として受理される関係記述の一例を示す。

正しい出力例

```
{  
    "sentence": "The apple is on the table.",  
    "object": ["apple", "table"],  
    "relation": "on",  
    "position": {  
        "vertical": "up",  
        "horizontal": "None",  
        "depth": "None"  
    }  
}
```

### 1.1.2 知識ネットワーク構築

本節では、前段で得られた関係データをもとに、物体、関係語、位置関係、および接触情報を統合して関係グラフを構築する。構築される知識ネットワークは、有向グラフ  $G = (V, E)$  で表され、ノード集合  $V$  は物体や関係に関する要素を、エッジ集合  $E$  はそれらの共起確率を表す。各ノードは、画像中の物体と LLM 出力の意味情報を対応づけ、関係文から抽出された主語・関係語・目的語を多段的に結ぶ構造を持つ。これにより、単なる文表現のつながりではなく、空間的配置と物理的整合性を伴う関係ネットワークを表現できる。構築する知識ネットワークの一例を図に示す。

(図)

#### ノード設計

本研究で用いるノード種別を表 1.1 に示す。これらのノードは、物体 → 位置付き物体 → 関係表現 → 三つ組という多層的接続を通じて結ばれており、言語的関係と空間的配置を統合的に扱うことが可能となっている。

表 1.1: ノード種類ごとの例と役割

ノード種別	例	役割
object	apple	画像内物体（主語・目的語）を表す基本ノード
area_object	up_apple	主語側の位置情報を付与したノード (up/down/left/right/front/back)
relation_object	apple on	主語と関係語を結合したノード
triplet	apple on table	完全な三つ組（主語・関係語・目的語）
relation_object_to	on table	関係語と目的語を結合したノード
area_object_to	down_table	目的語側に反転位置を付与したノード (逆方向探索用)

area\_object ノードは、言語的な関係表現が必ずしも一意な物理配置を示さないという問題を解決するために導入する。例えば、「The apple is on the table.」という文を考えると、人間は直感的に「リンゴが机の上にあり、上方向の接触が生じている」状況を理解できる。しかし同じ”on”には、「壁にかかった絵」や「ドアに貼られたメモ」のように、接触方向が上方向ではなく側面方向である場合にも用いられる。このとき、”on”を単独の関係語として集約し、単純に object ノード → relation\_object ノード (apple → apple on) のような表現だけでネットワークを作ると、異なる物理配置に基づく”on”が同一ノードに混在してしまう。その結果、探索時に参照される重みは「”up”だから”on”になった」「”front”だから”on”になった」といった位置条件を反映したものではなく、「単に”on”という単語がよく生成されたから」という頻度バイアスに引っ張られた結果になり得る。これでは、探索経路が表す意味が曖昧になり、なぜその関係を選んだのかを位置情報に基づいて説明できない。

そこで本研究では、位置情報をノードとして明示化し、up\_appleのような area\_object ノードを relation\_object ノードの前段に挿入する。この設計により、「上方向にあ

る”apple”からは”on”/”above”が候補になりやすい」といった位置条件付きの関係傾向がグラフ上に分離され、探索が単語頻度に支配されることを抑制できる。

さらに、目的語側にも area\\_object\\_to ノードを生成する。これは記述された位置情報の逆方向 (up↔down, right↔left, front↔back) に基づいて、主語側の位置属性と整合する形で目的語側の位置を表現するものである。これにより、探索は主語側だけに依存せず、目的語側からも一貫した制約で関係を評価できる。

## エッジ設計と重み付け

本研究における知識ネットワークの重み設計は、LLM が人間によって書かれた大量の文章を学習しており、常識的・経験的に妥当な関係性ほど高頻度で生成されるという性質に基づいている。物体ペアに対して関係文を大量に生成させ、その中で頻出する関係性ほど信頼度が高いと仮定し、出現頻度に基づいて知識ネットワークのエッジを定義する。

まず、object ノード →area\\_object ノードおよび object ノード →area\\_object\\_ノード対応するエッジは、物体から位置付き物体への遷移を表す構造的エッジである。これは、画像から得られる情報を探索に反映させることを目的としているため、エッジの重みは定数を与える。

$$w_{\text{object} \rightarrow \text{area}} = k \quad (k \text{ は定数}) \quad (1.1)$$

次に、area\\_object ノード →relation\\_object ノードに対応するエッジは、特定の位置条件にある物体がどの関係語を取りやすいかを表す。例えば up\\_apple →apple on は、「上方向に位置する apple が on という関係で表現される妥当性」を示す遷移である。よって、条件付き確率として重み付けを行う。具体的には、以下のように定義する：

$$\begin{aligned}
w_{\text{area\_object} \rightarrow \text{relation\_object}} &= P(\text{relation} \mid \text{area}, \text{object}) \\
&= \frac{\text{count}(\text{area}, \text{relation}, \text{object})}{\sum_{\text{relation}} \text{count}(\text{area}, \text{relation}, \text{object})}.
\end{aligned} \tag{1.2}$$

この定義により, ”on” という語が一般的に頻出から選ばれるのではなく, その位置条件において支持された関係語が高く評価される.  $\text{area\_object\_to}$  ノード  $\rightarrow \text{relation\_object\_to}$  ノードのエッジも同様に定義する.

$\text{relation\_object}$  ノード  $\rightarrow \text{triplet}$  ノード および  $\text{relation\_object\_to}$  ノード  $\rightarrow \text{triplet}$  ノードのエッジは, 特定の物体ペアに対して, どの関係語が成立しやすいかを表す. ある 2 つの物体が与えられた時, 同じ位置条件を満たしていたとしても, 成立しやすい関係語は物体の組み合わせによって異なるため, 物体ペアを条件とした関係成立のしやすさを条件付き確率として重み付けする. 具体的には, 次式によって定義する:

$$\begin{aligned}
w_{\text{relation\_object} \rightarrow \text{triplet}} &= P(\text{relation} \mid \text{object}_1, \text{object}_2) \\
&= \frac{\text{count}(\text{object}_1, \text{object}_2, \text{relation})}{\sum_{\text{relation}} \text{count}(\text{object}_1, \text{object}_2, \text{relation})}.
\end{aligned} \tag{1.3}$$

このように定義することで, 物体ペアに対してより適した表現の関係語を高く評価される.

## 接触関係の付与

1.1.1 で生成された”attach”は, 生成文が想定する状況における物理的接触の有無を表す. グラフ構築の際, この情報を triplet ノードの属性として付与する.

三つ組  $(o_1, r, o_2)$  に対して生成された文集合を  $\mathcal{S}_{(o_1, r, o_2)}$  とし, 各文  $s \in \mathcal{S}_{(o_1, r, o_2)}$  に付与された接触判定を  $\text{attach}(s) \in \{\text{True}, \text{False}\}$  とする.

このとき, True および False の出現回数をそれぞれ

$$\begin{aligned} N_{\text{True}} &= \sum_{s \in \mathcal{S}_{(o_1, r, o_2)}} \mathbb{I}[\text{attach}(s) = \text{True}], \\ N_{\text{False}} &= \sum_{s \in \mathcal{S}_{(o_1, r, o_2)}} \mathbb{I}[\text{attach}(s) = \text{False}] \end{aligned} \quad (1.4)$$

と定義する。ここで  $\mathbb{I}[\cdot]$  は指示関数である。

最終的な三つ組ノードに付与する接触属性  $\text{Attach}(o_1, r, o_2)$  は、次式により決定する：

$$\text{Attach}(o_1, r, o_2) = \begin{cases} \text{True} & \text{if } N_{\text{True}} > N_{\text{False}}, \\ \text{False} & \text{otherwise.} \end{cases} \quad (1.5)$$

これにより、接触関係を探索条件として利用でき、物理的整合性を損なう推論を抑制できる。

## 1.2 シーングラフ生成

本節では、入力画像に対してシーングラフを生成する手順を述べる。提案手法では、前節で構築したネットワークを固定の知識基盤として用い、画像から得られる観測情報を探索の手がかりとして関係推論を行う。具体的には、第一に物体認識により画像内の物体集合とその位置情報を推定し、第二に推定された物体ペアの相対位置・接触判定を探索条件として関係グラフを探索することで、各物体ペアに対する関係を決定する。得られた三つ組を統合することで、画像に対応するシーングラフを生成する。

### 1.2.1 物体認識

本手法では、まず入力画像から物体候補を列挙し、続いて各物体の位置情報を推定する。物体候補の列挙にはLLMを用い、画像内に存在しうる主要物体名をオープンボ

キャプラリに抽出する。次に、列挙された物体名をテキストクエリとして Grounding DINO に入力し、各物体のバウンディングボックスを推定することで、画像内の物体位置を得る。さらに、物体領域をより明確に抽出するため、SAM によるマスク生成を併用し、重なり合う物体や境界が曖昧な領域に対してもピクセルレベルでの物体形状を取得する。これらの処理により、後段の関係推論に必要となる物体集合と空間的手がかりを安定して得ることができる。

使用モデルについての説明は以下の通りである。

## GroundingDINO

Grounding DINO[?] は、Liu らによって提案されたオープンセット物体検出モデルである。従来の物体検出モデルは COCO などのデータセットに含まれる固定カテゴリしか検出できないという制約があった。これに対して、Transformer ベースの物体検出器である DINO を基盤とし言語情報を導入することで、自然言語の入力を条件として任意の物体を検出できるように設計されている。そのため、未知のカテゴリを含むオープンワールド環境でも柔軟に対応可能である。

Grounding DINO は、画像とテキストの情報を密接に統合するために、Dual-encoder-single-decoder アーキテクチャを採用している。従来モデルが一部の段階でしか特徴量を融合していなかったのに対し、Grounding DINO では以下の 3 つのフェーズで視覚と言語を融合させている。図 1.1 に Grounding DINO のフレームワークを示す。

## Dual-encoder-single-decoder アーキテクチャ

### 1. Feature Enhancer

自己注意機構に加え、テキストから画像、画像からテキストへの双方向の交差注意機構（Cross-Attention）を積み重ねることで、特徴量を強化・融合する。

### 2. Language-guided Query Selection

入力テキストに最も関連性の高い画像特徴をデコーダのクエリとして選択する。

### 3. Cross-modality Decoder

画像とテキストの両方の特徴量をクエリに注入し、モーダル間のアライメントをさらに高め、最終的な物体ボックスとラベル語句を出力する。

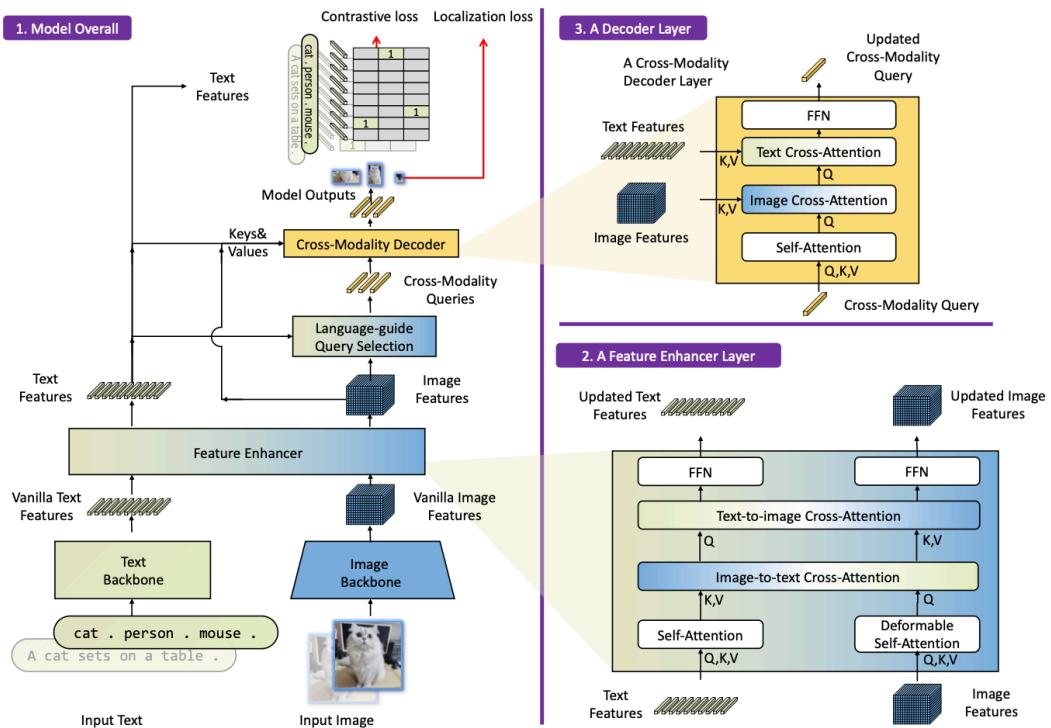


図 1.1: Grounding DINO のフレームワーク<sup>1</sup>

さらに、複数のカテゴリ名を入力する際、無関係な単語同士が干渉し合うのを防ぐため、Sub-sentence level Representation という技術を導入している。Attention mask を用いて、特定の単語間の関連性を強制的に遮断し、単語ごとの細かい情報

<sup>1</sup>[?] から引用

を保持している。学習には、物体検出データ（Objects365, OpenImages, COCO）、グラウンディングデータ（Flickr30k, Visual Genome, RefCOCOなど）、キャンプションデータ（Cap4M）を含む大規模データセットで事前学習を行うことで、高度な概念の一般化を実現している。

## SAM

以上2つのVLMモデルとGPT-4oを使用し、物体を認識させる。

具体的な手段としては、まず画像をLLMに入力し、画像内に含まれる可能性のある主要な物体を列挙する。これは未知の物体名を含むオープンボキャブラリな物体候補集合を得ることを目的とする。出力は修飾語を含まない単純名詞のリストであり、以下のプロンプトをLLM（GPT-4o）に与えることで行う。なお、実際のプロンプトは付録に示す。

### プロンプトの概要 —————

以下の指示、出力形式に基づいて、画像内の物体を抽出することを指示  
#指示

1. 画像内に見える**主要な物体**を全て列挙
2. 各要素は**単純な単数系の名詞**
3. **形容詞・色・素材・状態**などは含めない
4. **認識された物体名のみ**を出力する

#出力形式

カンマで区切る

大規模言語モデルは、事前学習によって膨大な知識や文脈理解能力を獲得しているが、その能力を効果的に引き出すためにはプロンプトの設計が極めて重要である。プロンプトはモデルに対する入力文であり、その内容や構造によって出力結果の精度・一貫性・創造性が大きく変化する。そこで上記のプロンプトでは、出力形式を明確に定義することでノイズを排除し、語彙の統一性を保つことで後段の処理を容易

にすることを目的として設計した。このような設計により、モデルが曖昧な表現や冗長な記述を避け、画像内の実体に対応する一般的な語彙を抽出できるようになる。

なお、ここで得られた物体名の集合を

$$\hat{O} = \{\hat{o}_1, \hat{o}_2, \dots, \hat{o}_n\} \quad (1.6)$$

とする。

次に、得られた物体名をテキストクエリとして Grounding DINO に入力し、対応するバウンディングボックス  $box = (x_1, y_1, x_2, y_2)$  と信頼度を得る。ここで、物体ごとの中心座標は次式で求める。

$$\mathbf{c}_i = \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \quad (1.7)$$

さらに、各物体領域をより明確に抽出するため、SAM（第 1.2.1 項）によるマスク生成を併用する。SAM は、Grounding DINO で得られた矩形領域を入力とし、物体境界をピクセルレベルで抽出する。これにより、矩形境界のみでは不明確だった物体の輪郭が精密に得られ、重なり合う物体や接触判定における曖昧性を低減することができる。SAM の出力は、COCO 形式のポリゴン表現  $segmentation = (x_0, y_0, x_1, y_1, \dots, x_n, y_n)$  で得る。これにより、物体  $o_i$  の位置情報として、矩形+ポリゴンのハイブリッド表現が得られる。

### 1.2.2 ネットワーク探索

本節では、入力画像から得られた 2 物体の座標情報（第項）を手がかりとして、事前に構築された知識ネットワーク内を探索し、物体間の意味的関係を推定する。ネットワーク探索に用いる入力情報は、物体認識段階で得られる物体名に加えて、2 物体間の相対位置関係および物理的接触関係である。これらは関係グラフ上の探索起点および探索制約として利用され、画像内の幾何構造と整合する関係推論を可能に

する。以下では、相対位置関係と接触関係の決定方法、およびそれらを用いた関係グラフ探索の手順について説明する。

## 相対位置関係の決定

2物体間の相対位置関係は、物体サイズを考慮した幾何学的判定に基づいて決定する。単純に中心座標のみを比較すると、大きさの異なる物体間では直感に反する判定が生じる場合があるため、本研究では物体サイズに応じて判定基準を切り替える。

まず、2物体の大きさが近しい場合には、物体の中心座標同士を比較し、垂直方向、水平方向、奥行方向の相対関係を決定する。この場合、小さい体同士であることが多く、中心点の相対位置がそのまま人間の直感に近い位置関係を表す。

一方で、物体サイズが大きく異なる場合、特に大きな物体と小さな物体の関係を推定する際には、中心座標のみを用いた判定は適切ではない。例えば、机の上に置かれたコップを考えると、机のバウンディングボックスは非常に大きいため、コップの中心点は机の左側や右側に位置していると判定される可能性がある。しかし、人間が重要視するのは「コップが机の上にある」という垂直方向の関係である。

この問題を避けるため、本研究では小さい物体が大きい物体の水平方向の幅に収まっているかどうかを確認し、収まっている場合には垂直方向の位置関係を重視して判定を行う。一方、小さい物体が大きい物体の高さに収まっている場合には、水平方向の関係を重視する。どちらにも該当しない場合には、中心座標に基づく判定を用いる。このような切り替えにより、重要でない方向の位置情報が探索に影響することを防ぎ、人間の直感に近い相対位置関係を抽出できる。

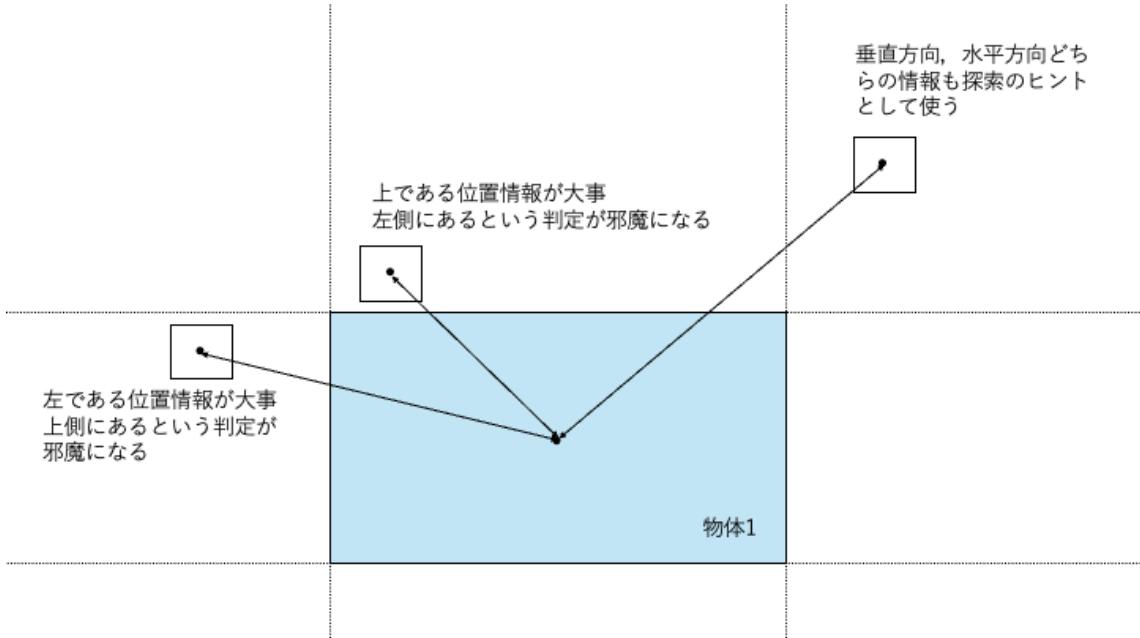


図 1.2: 位置関係の決定（仮）

## 接触関係の決定

次に、2 物体間の物理的接触関係を判定する。接触判定は計算効率と精度の両立を目的として、二段階の処理によって行う。

まず、Grounding DINO によって得られた2 物体のバウンディングボックスを用い、粗い接触判定を行う。バウンディングボックス同士が空間的に重なっていない場合、物体が接触している可能性は低いため、この時点で非接触と判定する。この処理は、後段の計算コストの高い処理を不要な場合に回避するための前処理として機能する。

次に、バウンディングボックスが重なっている場合のみ、SAM によって得られたセグメンテーションマスクを用いた精密な判定を行う。ただし、SAM のマスクは物体境界を厳密に抽出するため、実際には接触している場合であっても、マスク同士が完全に重ならないことがある。そこで本研究では、両物体のマスクを数ピクセル分膨張させた上で、それらが重なり合うかどうかを判定する。この処理により、境界付近の微小な誤差による誤判定を抑えつつ、物理的接触の有無を安定して判定す

ることが可能となる。

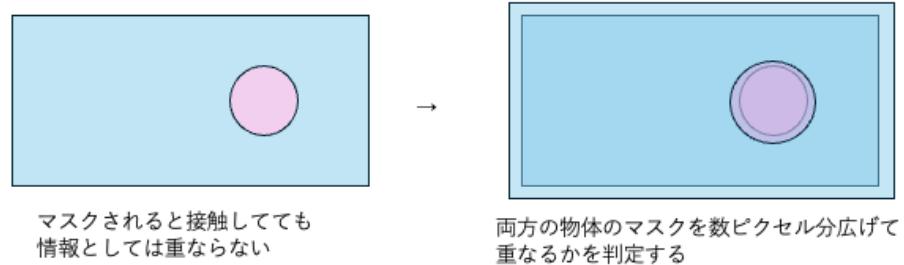


図 1.3: 接触関係の決定（仮）

### 関係グラフ探索と関係推論

以上の手順によって決定された相対位置関係および接触関係を用いて、関係グラフ上の探索を行う。

関係グラフの各エッジには、LLM によって生成された関係文の出現頻度に基づく確率的な重みが付与されている。本研究では、これらの確率を用いて複数エッジからなる経路の尤度を評価するため、エッジ重みをコストに変換し、最短経路探索として問題を定式化する。具体的には、確率の積を最大化する問題を、コストの和を最小化する問題へと変換することで、効率的な探索を可能にしている。

探索は、決定された相対位置関係に対応する位置付き物体ノードを始点として行う。主語側と目的語側の双方から探索を行い、両方の始点から到達可能な triplet ノードのうち、総コストが最小となるものを推定結果として採用する。この双方向探索により、一方の視点のみに依存した偏った推論を避け、より安定した関係推定が可能となる。

また、triplet ノードには接触属性が付与されているため、探索時には入力画像から得られた接触判定と一致するノードのみを候補として考慮する。これにより、空間的には近いが物理的に成立しない関係や、接触条件に矛盾する関係が推定されることを抑制できる。

さらに、相対位置関係が曖昧で複数の方向候補が存在する場合には、それぞれの位置条件を独立に用いて探索を行い、得られた結果のうち尤度が最も高いものを最終的な関係推論結果として採用する。この処理により、単一の位置判定に依存しない柔軟かつ頑健な関係推論が実現される。

### 1.2.3 シングラフ生成

前段のグラフ探索により得られた物体ペアごとの関係推論結果を統合し、最終的なシングラフを生成する。具体的には、各物体対  $(o_i, o_j)$  に対して推定された関係語を用いて、〈主語、関係、目的語〉からなる三つ組を構成する。これらの三つ組の集合を、物体をノード、関係をエッジとする構造として表現したものを、本研究におけるシングラフと定義する。