



Attribute Annotation Template:

[category, color, shape, size, material, surface texture, position, state, function, brand/model, interactivity, proximity to the person]

```
{
  "object": "Backpack",
  "attributes": {
    "Category": "個人アイテム",
    "Color": "緑 (主要色)、黒 (副色)",
    "Shape": "角が丸い長方形",
    "Size": "中型",
    "Material": "布",
    "Surface Texture": "滑らか",
    "Position": "ソファの上",
    "State": "使用済み",
    "Functionality": "収納、持ち運び",
    "Brand/Model": "指定なし",
    "Interactivity": "静的、可搬",
    "Proximity to Person": "近い、ソファの左アームの上"
  }
}
```



```
{
  "object": "Laptop",
  "attributes": {
    "Category": "電子機器",
    "Type": "ノートパソコン",
    "Color": "黒",
    "Shape": "長方形",
    "Size": "中型",
    "Material": "プラスチック、金属",
    "Surface Texture": "滑らか",
    "Position": "壁際の床の上",
    "State": "使用済み、画面オン",
    "Functionality": "計算、情報表示",
    "Brand/Model": "指定なし",
    "Interactivity": "インタラクティブ、有線接続",
    "Proximity to Person": "遠い、壁際の床の上"
  }
}
```

Model Inputs

Images: I



Dialogue Texts:
 $D = \{d1, d2, \dots, dn\}$

Human utterance:

“ティッシュボックスを緑のバックパックの中に入れる”

Prompt:

[System]
[User]
[Output]

Multimodal End-to-End Perception (f_{VLM}, θ)

Object Identification

Human utterance:

“ティッシュボックスを緑のバックパックの中に入れる”

$O = \{o1, o2, \dots, om\}$

Identified objects:
ティッシュボックス & 緑のバックパック

$\lambda_{obj} \cdot \mathcal{L}_{obj}$

Text-to-Image Object Reference



“緑のバックパック” & “ティッシュボックス”

$\lambda_{ref} \cdot \mathcal{L}_{ref}$

Next Action Prediction



Robot arm go to tissue box?

Action:
テーブル上のティッシュボックスを掴む

$\lambda_{act} \cdot \mathcal{L}_{act}$

$$\mathcal{L}_{joint} = (\lambda_{obj} \cdot \mathcal{L}_{obj}) + (\lambda_{ref} \cdot \mathcal{L}_{ref}) + (\lambda_{act} \cdot \mathcal{L}_{act})$$