

# Failure Taxonomy 标注指南

---

## 标注目标

对每条 **Failure@1** (GT 不在 Top-1 ) 进行**单标签**归类：

- 只选 一个**最主要原因** (primary failure mode) : Ambiguous (underspecified, nearduplicate); Attribute; Action; Count; Context; Sptial; Object;
- 若 10 秒内无法稳定归类 → 选 **Ambiguous**
- 标注对象是：**caption (query) vs 左图 GT vs 右图 Top-1**
- 每人标注：overlap + ABC中选一个，如果有图片损坏，从backup中抽一个替补。

## Notice

标注时注意对应idx和img的idx，不要错位；可以提前对csv进行排序和img的idx顺序（升序）一致。

---

## 总规则（必须遵守）

---

### R0. 单标签原则

- 每条只填一个 **category**
- 允许在 **ambiguous\_subtype** 里补充 Ambiguous 的子类（可选）

### R1. 10 秒规则

- 如果你在 10 秒内在两个类别间摇摆 → **Ambiguous**

### R2. “Top-1 也满足 caption”就不要硬凑语义错误

- 如果 GT 和 Top-1 都合理满足 **caption** (尤其 COCO caption 很泛) → **Ambiguous**

### R3. Object Confusion 的判别句

问自己一句：模型把东西认成“另一个类别/漏掉必须出现的物体”了吗？

- 是 → Object Confusion
  - 否 → 多半是 Attribute / Action / Spatial / Ambiguous
- 

## 类别定义（7 类）

---

### 1) Attribute Binding (属性/状态/局部细节绑定错误)

定义

Top-1 找到了大致对象/场景，但没满足 **caption** 中明确的属性或状态（颜色、材质、部件存在、是否被咬、是否打开等）。

### 何时使用

- 颜色/材质/品牌/衣着等属性错
- “bitten / broken / open / closed / dirty / clean / sliced”等状态错
- 部件存在性（hand, wheel, handle, logo 等）若 caption 明确要求

### 不要误用

- 若对象类别本身错（zebra→horse）→ Object Confusion
- 若 caption 没提到该属性 → 可能 Ambiguous

### 例子

- caption: “a **bitten** hotdog” → pred: hotdog 完整
- caption: “a **red** bus” → pred: blue bus
- caption: “pizza with **pineapple**” → pred: 无 pineapple

---

## 2) Object Confusion ( 类别错 / 缺失必须对象 )

### 定义

Top-1 在“必须出现的名词对象”上失败：

- (a) 把对象认成别的类别
- (b) 漏掉 caption 明确要求的一个对象（missing object）

### 何时使用

- zebra ↔ horse, snowboard ↔ skis
- caption 明确列出两个物体：A and B，但 pred 少了 B
  - “a sink and a toilet” → pred 只有 sink（**missing object**）

### 不要误用

- 只是角度/光照差异 → Ambiguous
- caption 没明确要求那个对象 → Ambiguous 或 Scene/Context

### 例子

- caption: “a **sink and a toilet**” → pred: 无 toilet
- caption: “two **zebras**” → pred: horses

---

## 3) Spatial Relation ( 空间关系错误 )

## 定义

caption 明确描述**相对位置/方向/几何关系**， Top-1 未满足：

- left/right, above/below, in front of/behind, on top of/under, next to/between

## 何时使用

- caption 对空间关系是“区分点”

## 不要误用

- caption 没写空间关系 → 不是 Spatial
- 空间差异太弱或你肉眼也不确定 → Ambiguous

## 例子

- caption: “dog **to the left of** a boy” → pred: dog 在右边
  - caption: “cup **on top of** plate” → pred: cup 在旁边
- 

## 4) Action / Interaction ( 动作/交互错误 )

### 定义

caption 以动词为核心 ( holding, riding, cutting, biting, throwing, playing ) ， Top-1 的动作/交互不匹配。

### 何时使用

- “person **holding** a phone” → pred: phone 在桌上
- “man **riding** a bike” → pred: bike 停着无人骑

### 不要误用

- caption 只是泛泛的 “a person with ...” 但动作不清晰 → 多半 Ambiguous
- 动作是对的，只是场景/对象细节差 → Attribute 或 Ambiguous

### 例子

- caption: “a person **cutting** a cake” → pred: 只是摆着蛋糕
- 

## 5) Scene / Context ( 场景/语境错 )

### 定义

caption 强调场景类别 ( kitchen, bathroom, beach, street, indoor station ) ， Top-1 场景不符合。

### 何时使用

- caption 的区分点是“在哪里发生”

- pred 与 GT 的场景类别明显不同

### 不要误用

- 两张都算同一大类场景但细节不同 → Ambiguous
- caption 太泛 ( indoor/outdoor不清 ) → Ambiguous

### 例子

- caption: "in a **kitchen**" → pred: 户外烧烤场景
  - caption: "train at an **indoor platform**" → pred: 明显户外铁轨 ( 若确凿 )
- 

## 6) Counting / Plurality ( 数量/单复数错误 )

### 定义

caption 明确数量/复数 ( two, three, several, many ) , Top-1 数量不符合。

### 何时使用

- caption 明确出现数量词或强复数线索

### 不要误用

- caption 没写数量词 → 不要因画面多/少就强判
- 数量难判断 ( 遮挡/远景 ) → Ambiguous

### 例子

- caption: "**two** dogs" → pred: 1 dog
  - caption: "**three** people" → pred: 1 person
- 

## 7) Ambiguous / Under-specified ( 不可判别/弱描述 )

### 定义

GT 与 Top-1 都合理满足 **caption** , 或 caption 信息不足以区分 , 或差异主要是低层视觉细节 ( 角度、光照、构图 ) 。

### 子类 ( 可选 , 建议填 )

- **underspecified**: caption 太泛 ( "a person in a kitchen" )
- **nearduplicate**: 多张图都满足 caption, Top-1 与 GT 近重复

### 使用规则 ( 强制 )

- 10 秒内无法稳定归类 → Ambiguous
- 肉眼都说不清对错 → Ambiguous

## 例子

- caption: "a person snowboarding on snow"  
GT 与 pred 都是人滑雪板 → Ambiguous
- caption: "people preparing food in a kitchen"  
两张都满足 → Ambiguous

---

## 实操注意事项 ( 避免团队不一致 )

---

### N1. "Missing object" 算 Object Confusion ( 统一口径 )

- "sink and toilet" 缺 toilet → **Object Confusion**
- 不要放到 Attribute

### N2. "bitten / open / broken / with/without hand" 统一算 Attribute Binding

- 除非 caption 的核心是动作 ( biting/holding ) 才算 Action/Interaction

### N3. 先判断 Ambiguous , 再判断类别

你们的最大时间浪费来自"强行归类"。顺序应当是：

1. 是否 Ambiguous ?
2. 否 → 再分 Attribute/Object/Spatial/Action/Scene/Counting

---

## 建议的团队流程 ( 对齐 + 可靠性 )

---

1. 全员先标同一份 `assign_overlap.csv` (30 条)
2. 对比冲突样本 , 更新 guide (v2)
3. 再标各自 50 条 unique
4. 最后合并统计并画图