Failure Taxonomy Annotation Guide

(Vision–Language Retrieval, MS-COCO, CLIP)

Purpose

This guide defines a **shared taxonomy and annotation standard** for labeling high-confidence retrieval failures in image–text retrieval experiments.
The goal is to ensure **consistency across annotators**, **reproducibility of analysis**, and **clarity in reporting**.

Annotations are performed on **Top-1 retrieval failures** only.

General Annotation Rules

R1. Single-label rule

Each failure must be assigned **exactly one primary category** representing the **main cause** of the failure.

R2. Ten-second rule

If a failure cannot be confidently assigned to a single category within **10 seconds**, label it as **Ambiguous**.

R3. No forced interpretation

If the retrieved image reasonably satisfies the caption, **do not force a semantic error**.
Such cases should be labeled **Ambiguous**.

R4. Caption-driven judgment

Annotations are based on the **caption semantics**, not on personal expectations of what “should” be shown.

R5. Categories and tasks:

- Only one primary failure mode:
  - Ambiguous (underspecified, nearduplicate);
  - **Attribute**;
  - **Action**;
  - **Count**;
  - **Context**;
  - **Sptial**;
  - **Object**.
- Further instructions are given later.
- Task: labelling: overlap + A/B/C. The backup is in case there is damaged image in sets, in which you can replace the orginal bad image with random one in the backup set.

Failure Categories

1. Attribute Binding

Definition:
The retrieved image contains the correct object or scene but **fails to satisfy an explicit attribute, state, or part-level detail** specified in the caption.

Typical attributes:

- Color, material, texture
- Object state (open/closed, bitten/unbitten, broken/intact)
- Presence or absence of a specific part (hand, wheel, logo)

Use this category when:

- The object category is correct, but an attribute or state is wrong.

Do NOT use when:

- The object category itself is incorrect (see Object Confusion).
- The attribute is not explicitly mentioned in the caption.

Examples:

- Caption: “a **bitten** hotdog” → Retrieved: unbitten hotdog
- Caption: “pizza with **pineapple**” → Retrieved: pizza without pineapple
- Caption: “a person with a **visible hand**” → Retrieved: no visible hand

---

2. Object Confusion

Definition:
The retrieved image fails to satisfy **mandatory object-level requirements** in the caption.

This includes:

- **Wrong object category**
- **Missing object** when the caption explicitly requires multiple objects

Use this category when:

- A required object is absent.
- An object is replaced by a different category.

Do NOT use when:

- Differences are minor or due to ambiguity.
- The caption does not strictly require the missing object.

Examples:

- Caption: “a **sink and a toilet**” → Retrieved: sink only
- Caption: “two **zebras**” → Retrieved: horses
- Caption: “a **snowboard**” → Retrieved: skis

3. Spatial Relation

Definition:
The retrieved image violates an **explicit spatial relationship** described in the caption.

Common relations:

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

Use this category when:

- The caption clearly specifies a spatial relationship that is incorrect in the retrieved image.

Do NOT use when:

- Spatial relations are not mentioned.
- Spatial differences are subtle or unclear.

Examples:

- Caption: “a dog **to the left of** a boy” → Retrieved: dog on the right
- Caption: “a cup **on top of** a plate” → Retrieved: cup beside the plate

4. Action / Interaction

Definition:
The retrieved image does not satisfy an **explicit action or interaction** described in the caption.

Typical actions:

- holding, riding, cutting, biting, throwing, playing

Use this category when:

- The caption centers on an action or interaction.
- The retrieved image depicts a different or absent action.

Do NOT use when:

- The caption does not clearly specify an action.
- The action is correct but other details differ.

Examples:

- Caption: “a person **holding** a phone” → Retrieved: phone on table
- Caption: “a man **riding** a bike” → Retrieved: bike unattended

5. Scene / Context

Definition:
The retrieved image does not match the **scene or environmental context** explicitly specified in the caption.

Common contexts:

- kitchen, bathroom, beach, street
- indoor / outdoor

Use this category when:

- The scene type is central to the caption and clearly violated.

Do NOT use when:

- The caption is vague about the scene.
- Both images belong to the same broad scene category.

Examples:

- Caption: “in a **kitchen**” → Retrieved: outdoor picnic
- Caption: “train at an **indoor platform**” → Retrieved: outdoor tracks

6. Counting / Plurality

Definition:
The retrieved image violates an **explicit numerical requirement** in the caption.

Use this category when:

- The caption includes a clear quantity or plurality cue.

Do NOT use when:

- No explicit number is given.
- Quantity is difficult to determine due to occlusion or scale.

Examples:

- Caption: “**two** dogs” → Retrieved: one dog
- Caption: “**three** people” → Retrieved: one person

7. Ambiguous / Under-specified

Definition:
The caption does not provide enough information to clearly distinguish between the ground-truth image and the retrieved image, or both reasonably satisfy the caption.

Subtypes (optional):

- `underspecified`: caption lacks distinguishing cues
- `nearduplicate`: multiple visually similar images satisfy the caption

Use this category when:

- Both images reasonably match the caption.
- Differences are low-level (angle, lighting, composition).
- You cannot confidently assign another category.

Mandatory rule:
If uncertain between two categories → **Ambiguous**.

Examples:

- Caption: “people preparing food in a kitchen” → Both images satisfy
- Caption: “a person snowboarding on snow” → Multiple valid matches

---

Annotation Workflow (Recommended)

1. Check if the caption provides **clear distinguishing cues**.
   - If not → Ambiguous.
2. If yes, identify the **single dominant failure cause**.
3. Assign exactly **one category**.
4. Do not over-interpret minor visual differences.

Notes on Consistency

- Missing required objects → **Object Confusion**
- Wrong state / part presence → **Attribute Binding**
- Unclear or reasonable alternative matches → **Ambiguous**

Intended Use

This taxonomy supports:

- Quantitative failure distribution analysis
- Qualitative case studies
- Failure-driven improvement experiments
- Transparent reporting of dataset ambiguity