Failure Definition and Taxonomy

This document defines the failure criteria and taxonomy used in the failure mode analysis of CLIP zero-shot retrieval.

Failure Definition

We define a **Failure@1** case as:

> A caption whose corresponding ground-truth image is **not ranked within the Top-1** retrieved image for that caption.

All failure analyses in this project are conducted under this definition.

Failure Taxonomy

Each failure case is assigned to exactly **one** of the following categories.

1. Spatial Relation

Failures involving incorrect understanding of spatial relationships
(e.g., left/right, front/behind, above/below).

Example:
"A dog on the left of a boy."

---

2. Attribute Binding

Failures where attributes such as color, size, or material are not correctly bound to the target object.

Example:
"A red bus parked on the street."

---

3. Object Confusion

Failures where visually similar object categories are confused.

Example:
"Zebra" vs. "horse", "ham" vs. "pizza".

---

4. Action / Interaction

Failures involving misunderstanding of actions or interactions between entities.

Example:
"A person riding a skateboard."

---

5. Scene / Context

Failures where the overall scene or environment is misinterpreted.

Example:
"A small kitchen with white cabinets."

---

6. Counting / Plurality

Failures involving incorrect understanding of quantity or plurality.

Example:
"Two dogs playing in a park."

Annotation Protocol

- Two annotators independently label a subset of failure cases.
- Disagreements are resolved by discussion using the above definitions.
- No new categories are introduced after taxonomy freezing.