Failure Taxonomy Summary (Annotated Retrieval Failures)

We annotated a subset of CLIP retrieval failures (R@1 failures) and grouped them into error categories. On the merged A/B/C annotations (excluding overlap), the distribution is:

- **Ambiguous**: 31.9%
- **Object**: 29.52%
- **Attribute**: 11.9%
- **Action**: 11.43%
- **Context**: 7.62%
- **Count**: 7.14%
- **Spatial**: 0.48%

This indicates that a substantial fraction of failures are **Ambiguous** (underspecified captions or near-duplicate images), which are not easily "fixable" via simple text-side heuristics. Among actionable categories, **Object** dominates, followed by **Attribute** and **Action**. This motivates evaluating targeted improvements on these actionable subsets.

Improvement Method: Category-Aware Prompt Ensembling

We evaluate a text-side improvement method that does **not** retrain CLIP: **prompt ensembling** on the query side.

Category-aware templates

Instead of using a single generic template set, we generate templates conditioned on the failure category being evaluated (e.g., Object-focused vs Attribute-focused templates). For a caption *c*, we produce a set of K templated prompts (K depends on the selected categories; e.g., K=10 for single-category tests and K=15 for Object+Attribute tests), encode them with CLIP's text encoder, and compute similarity to all image embeddings.

Pooling across templates

Given K similarity score vectors for one caption, we aggregate them into a single similarity vector using one of:

- **max pooling**: take the maximum similarity per image over templates
- **mean pooling**: average similarity per image over templates
- **logsumexp pooling** (soft-max pooling): $(\tau \log \sum_{k=1}^K \exp(s_k/\tau))$ with $(\tau = 1.0)$

We then compute retrieval metrics R@1/R@5/R@10 on the annotated subset indices.

Results: Effect of Pooling Strategy by Category

Below we report baseline vs improved performance and the change in percentage points (pp). All runs use the same cached embeddings and seed=42.

1) Object + Attribute subset (n=87, K=15)

Baseline: **R@5 = 36.78%**, **R@10 = 51.72%**

Pooling	Improved R@5	ΔR@5 (pp)	Improved R@10	ΔR@10 (pp)
max	40.23%	+3.45	52.87%	+1.15
mean	**43.68%**	**+6.90**	52.87%	+1.15
logsumexp (τ=1.0)	**43.68%**	**+6.90**	52.87%	+1.15

Interpretation. For the dominant actionable subset (Object+Attribute), **mean/logsumexp pooling doubles the R@5 gain** relative to max. With larger template sets (K=15), max pooling is more sensitive to a single noisy/high-scoring template, while mean/logsumexp better reflects consistent support across templates.

2) Object subset (n=62, K=10)

Baseline: **R@5 = 33.87%**, **R@10 = 51.61%**

Pooling	Improved R@5	ΔR@5 (pp)	Improved R@10	ΔR@10 (pp)
max	37.10%	+3.23	53.23%	+1.61
mean	37.10%	+3.23	54.84%	+3.23
logsumexp (τ=1.0)	**38.71%**	**+4.84**	**54.84%**	**+3.23**

Interpretation. Object errors benefit consistently from prompt ensembling, and **logsumexp pooling yields the best overall gains** on both R@5 and R@10.

3) Attribute subset (n=25, K=10)

Baseline: **R@5 = 44.00%**, **R@10 = 52.00%**

Pooling	Improved R@5	ΔR@5 (pp)	Improved R@10	ΔR@10 (pp)
max	44.00%	+0.00	52.00%	+0.00
mean	**48.00%**	**+4.00**	52.00%	+0.00
logsumexp (τ=1.0)	**48.00%**	**+4.00**	52.00%	+0.00

Interpretation. Attribute failures show limited improvement overall. Gains appear mainly at **R@5** under mean/logsumexp pooling, while **R@10 remains unchanged**, suggesting that many attribute-related confusions are fine-grained and not easily resolved by text prompt variation alone.

4) Action subset (n=24, K=10)

Baseline: **R@5 = 33.33%**, **R@10 = 45.83%**

Pooling	Improved R@5	ΔR@5 (pp)	Improved R@10	ΔR@10 (pp)
max	37.50%	+4.17	66.67%	+20.83
mean	33.33%	+0.00	66.67%	+20.83
logsumexp (τ=1.0)	33.33%	+0.00	66.67%	+20.83

Interpretation. Action failures exhibit a large improvement at **R@10** across pooling strategies, indicating that prompt ensembling can push correct images into the top-10 results even when top-5 improvements are limited. However, the Action subset is small (n=24), so variance is expected; this effect should be interpreted cautiously.

Overall Findings

1. **Taxonomy reveals large Ambiguous fraction (31.9%)**: a substantial portion of failures are not clearly actionable with prompt-based methods.
2. **Prompt ensembling improves actionable failures**, especially **Object-heavy** subsets:
    - Object+Attribute: best ΔR@5 = **+6.90pp** (mean/logsumexp)
    - Object: best ΔR@5 = **+4.84pp** (logsumexp)
3. **Pooling strategy matters**:
    - For larger K and mixed categories, **mean/logsumexp pooling** is significantly better than max for R@5.
    - A single robust default is **logsumexp pooling (τ=1.0)**, which performs best or tied-best across the main actionable subsets.
4. **Attribute failures are harder**: improvements are smaller and mostly confined to R@5, suggesting limitations of text-only prompt variation for fine-grained visual distinctions.

Recommended "Improved" Pipeline Setting

Based on the ablation results, we recommend the following default improvement setting for further experiments:

- **Category-aware templates**
- **Pooling: logsumexp**, **τ = 1.0**
- Evaluate primarily on **actionable categories** (excluding Ambiguous)

This setting yields the strongest and most consistent improvements across major actionable categories without retraining the CLIP model.

Limitations

- Some category subsets are small (e.g., Action n=24, Attribute n=25), so observed gains may have high variance.
- Ambiguous failures (underspecified or near-duplicate cases) likely require additional information or different modeling approaches beyond prompt ensembling.