Research Improvement Plan: SPIN + MOD Optimization

1. Layer-Specific Application

- **Action:** Restrict SPIN application to the **last 50% of layers** (e.g., layers 16–32) instead of all layers (0–32).
- **Reason:** Early transformer layers typically handle syntax and grammar (text-heavy), while later layers handle semantic reasoning (image-heavy). Suppressing early layers disrupts basic language fluency.
- **Expected Result:** Improved overall fluency and sentence structure; recovery of "Attribute" scores.

2. Digit-Aware Alpha (Fixing the Counting Paradox)

- **Action:** Implement a conditional check in the decoding loop: `if token in [0-9] or token_type == 'number', set alpha_contrastive = 0.0`.
- **Reason:** Text priors are often *correct* for small numbers (e.g., "two eyes"). Standard contrastive decoding penalizes these high-probability correct guesses, forcing the model to pick a wrong number.
- **Expected Result:** "Counting" scores will recover to near-baseline levels while maintaining "Environment" improvements.

3. Threshold Tuning

- **Action:** Increase `js_threshold` from `0.08` to `0.12`.
- **Reason:** A low threshold triggers Contrastive Decoding too frequently, even for safe, grounded tokens. This "over-correction" degrades performance on simple tasks like Comparisons.
- **Expected Result:** Reduced "false positives" in hallucination detection; stability in "Comparison" and "Other" categories.

4. Aggressive Late-Stage Suppression

- **Action:** If implementing Step 1 (Layer Restriction), decrease `keep_head_ratio` from `0.95` to `0.80` *only for those later layers*.
- **Reason:** Once you stop damaging the early syntax layers, you can afford to be more aggressive in the reasoning layers to force stronger image grounding.
- **Expected Result:** Significantly higher scores in "Relation" and "Environment" due to forced visual reliance.

5. Consensus Guardrail

- **Action:** Add a check: `if top_token(Vision) == top_token(Text), use Greedy Decoding`.
- **Reason:** If both the vision model and the text prior agree on the same word, it is likely the correct, obvious answer. Subtracting logits in this scenario only introduces noise.
- **Expected Result:** General stability improvement across all metrics; prevents the model from over-thinking simple questions.