

Agentic Prompt Optimization with Visual Contrastive Reasoning for Fine-Grained Classification

Supplementary Material

A. APO-Pair Working Example

This section presents a complete working example of APO-Pair applied to fine-grained botanical classification, demonstrating the system’s ability to evolve sophisticated domain expertise through iterative prompt optimization. The task involves distinguishing between visually similar lupine species (*Lupinus arcticus*, *L. argenteus*, and *L. arizonicus*) based on subtle morphological differences, flower characteristics, and habitat context. Refer to Table 5 for detailed overview.

B. Hyperparameters

B.1. Optimization Parameters

- **Maximum iterations:** $T_{\max} = 60$
- **Early stopping:** No improvement limit = 10 iterations
- **Thompson sampling priors:** $\alpha_0 = \beta_0 = 1$ (uniform Beta prior)

B.2. Balanced Feedback Configuration

- **Feedback strategy:** Balanced equal representation
- **Maximum error examples:** 2 per iteration
- **Maximum correct examples:** 2 per iteration
- **Include correct examples:** True
- **Maximum images per call:** 8 (4 pairs due to Claude 3.7 Sonnet 10 images limit)

B.3. Agent Configuration

- **Model:** Claude 3.7 Sonnet
- **Temperature:** 0.1

B.4. Context Management

- **Rejected patches memory:** 20 most recent patches
- **Examples rotation:** Diverse selection strategy to prevent overfitting

B.5. YAML Patch Processing

- **Patch format:** Structured YAML with $+/ \sim /-$ operators
- **Auto-correction:** Automatic quote balancing for malformed YAML
- **Key format:** String keys without quotes (e.g., "1", "2", "3")
- **Maximum rule length:** 30 words per guideline (enforced via agent instructions)
- **Eval VLM:** Qwen-2VL-7B

C. Base Prompt

You are an image analyzer specializing in pairwise comparison.

BEGIN_EVAL_GUIDELINES

1: "Compare the two images"

END_EVAL_GUIDELINES

****Output Format Requirements**:**

- Give a concise evaluation reasoning comparing the images based on the guidelines
- Provide your final answer as either "same" (same family) or "different" (different family)
- Be precise and confident in your assessment

Provide response in this exact JSON format, with no additional text and no extra whitespaces. No extra comment, no prose:

```
{
  "evaluation": "Detailed comparison analysis of the images (30 words max)",
  "answer": "same" or "different"
}
```

D. Agent Instructions

Below are the complete instructions provided to the multimodal LLM agent (Claude 3.7 Sonnet) for proposing YAML patches during prompt optimization.

You are the Prompt Evolution Agent. Your goal is to improve the EVAL block between ### BEGIN_EVAL_GUIDELINES ... ### END_EVAL_GUIDELINES so that future predictions improve F1 score and accuracy for a pairwise images comparison task.

The task is to determine whether two images show the same species or different species. Create rules that are evaluation guidelines that would be ideal for comparing two images and distinguishing whether they are the

Table 5. Working example of APO-Pair optimization for lupine species classification. Round 1 shows the optimization steps in more detail (green) with Thompson sampling, context aggregation, agent analysis, YAML patch generation, new prompt scoring, and logging. Subsequent rounds show condensed results for brevity.

Iteration	Component	Content
Round 1	Initial Evaluation Base Prompt	You are an image analyzer specializing in pairwise comparison. ### BEGIN_EVAL_GUIDELINES 1: "Compare the two images" ### END_EVAL_GUIDELINES
	Score: 0.57	
	Log Info Context DB	Log prompt version into DB along with its score.
	Sampling Bandit selection process	Selected: Base prompt (v1) - unexplored arm with high potential Beta posterior: $\alpha = 1, \beta = 1$ (uniform prior) Strategy: Exploration of new prompt variant
	Context Aggregation	Balanced feedback from training set (P): Error examples: 2 misclassified pairs (false positives); Correct examples: 2 correctly classified pairs Rotation mechanism: Fresh examples to prevent overfitting
	Agent Analysis	Claude 3.7 Sonnet receive as input: The balanced feedback sample examples, current prompt, and a list of previously rejected patches if any. Agent is instructed to suggest improvements for the content in between: ### BEGIN_EVAL_GUIDELINES ... ### END_EVAL_GUIDELINES Instructions include: Think step-by-step, with deep reasoning process, but **only output** the final YAML patch (or exactly '# no_change')
	YAML Patch Generation and Merging	~1: "Compare the two plant images focusing on specific morphological features: flower structure, color patterns, leaf arrangement, growth habit, and overall plant architecture." +2: "Examine flower shape, size, and color." +3: "Examine leaf structure including number of leaflets, their arrangement, and texture as these are key species identifiers."
	New Prompt Scoring and Logging	Evaluate $\hat{P}(t)$ on P_{val} with evaluation VLM model (Qwen-2VL-7B) ✅ Accepted → v2 (0.584). Insert $\hat{P}(t)$ into R and update Beta counts; Log predictions to context DB
Round 2	1. Sampling Process 2. Context Aggregation 3. Agent Analysis 4. YAML Patch Generation and Merging	Bandit picked v2 🎲 ($\theta=0.89$). Retrieved cached score (0.584) Generate feedback bundle: 2 errors, 2 correct examples and List of rejected patches. <u>Patch Suggestion:</u> ~1: "Compare leaf structure including shape, size, number of leaflets, and texture, as different lupine species have distinctive leaf characteristics." ~2: "Analyze flower characteristics including color intensity, petal shape, keel structure, and banner patterns which vary between lupine species." +4: "Examine flower spike density, length, and arrangement, as species differ in inflorescence structure and flower spacing."
	New Prompt Scoring and Logging	Evaluate $\hat{P}(t)$ on P_{val} with evaluation VLM model (Qwen-2VL-7B) ❌ No improvement (0.580). Log rejected patches logged for future avoidance.
	Rounds keep running with this same setup	
Round 58	1. Sampling Process 2. Context Aggregation 3. Agent Analysis 4. YAML Patch Generation and Merging	Bandit picked v4 🎲 ($\theta=0.93$). Retrieved cached score (0.624) Generate feedback bundle: 2 errors, 2 correct examples and List of rejected patches. <u>Patch Suggestion:</u> ~2: "Analyze flower color precisely - Arctic lupine has blue-purple flowers, Silvery has pale blue, Arizona has bright pink-purple/magenta flowers that are distinctly more vibrant." ~8: "Compare habitat context closely - Arctic prefers alpine tundra with yellow ground cover, Silvery favors mountain meadows, Arizona thrives in dry desert conditions with sandy soil." +12: "Note soil conditions - Arctic lupines grow in moist alpine environments, while Arizona lupines prefer dry, sandy desert soils." +13: "Examine background vegetation as habitat indicators - yellow flowering ground cover often accompanies Arctic lupine in tundra environments."
	New Prompt Scoring and Logging	Evaluate $\hat{P}(t)$ on P_{val} with evaluation VLM model (Qwen-2VL-7B) ✅ Accepted → v17 (0.644). Insert $\hat{P}(t)$ into R and update Beta counts; Log predictions to context DB

same species or different species. You can update, remove any of the guidelines and you can add new guidelines as well.

Rules

1. Think step-by-step, with deep reasoning process, but ****only output**** the final YAML patch (or exactly '# no_change'). No extra commentary, no prose.
2. Keys in the YAML block are **strings of integers** ('1', '2'...). Use them exactly-e.g. '~ 5: "new text"' to modify rule 5.
3. Prefer **minimal and concise** edits:
+ #: "<new rule text>" (make this concise and all in a single line, 30 words max direct to the point)
- #: "<text to remove from SOP>"
~ #: "<new text>" (do NOT include the old text).
4. Fix the supplied errors ****without**** breaking previous correct cases. Make sure every quoted value ends with a closing quote.

5. Focus on comparative analysis between two , not image identification.
6. If no PROMPT change can fix the errors, output exactly: # no_change

Current PROMPT

```
```yaml
{current_sop}
```
```

#####

Previously rejected guidelines patch snippets (Use this list to gain context of what was tried before and DO NOT repeat or suggest similar patches, ONLY suggest new patches or guidelines, think of novel perspectives)

```
{blacklist}
```

#####

Feedback Bundle: This contains both ERRORS (misclassified pairs) and CORRECT examples (successfully classified pairs) to provide balanced learning signals.

```

### ERRORS - These pairs were
misclassified and need to be fixed:
Each error shows:
- ground_truth: The correct answer (Same
species or Different species)
- model_verdict: What the model
incorrectly predicted (Same species or
Different species)
- model_reasoning: The model's reasoning
evaluation for its wrong decision
- Images: Two photos that were compared

### CORRECT EXAMPLES - These pairs were
classified correctly and should be
preserved:
Each correct example shows:
- ground_truth: The correct answer (Same
species or Different species)
- model_verdict: What the model correctly
predicted (matches ground_truth)
- model_reasoning: The model's reasoning
evaluation for its correct decision
- Images: Two photos that were compared

**IMPORTANT**: Your improvements should
fix the errors while maintaining the
reasoning patterns that led to correct
classifications.

Here are the examples:
"""
{examples organized here}

##### CRITICAL OUTPUT FORMAT #####
You MUST output ONLY the YAML patch in
this exact format:

For adding new rules:
+ 2: "New rule text here"
+ 3: "Another new rule"

For modifying existing rules:
~ 1: "Modified rule text"

For removing rules:
- 2: "Rule text to remove"

For no changes:
# no_change

RULES:
- NO prose, explanations, or commentary
- NO markdown code blocks
- NO extra text before or after the patch
- Each rule must be on ONE line, max 30
words
- All quoted text must have closing quotes

```

- Keys are integers: 1, 2, 3, etc.

OUTPUT ONLY THE YAML PATCH NOW:

E. Full Optimized Prompts

The prompts below show only the evolved evaluation guidelines section that APO-Pair modifies during optimization (content between `### BEGIN_EVAL_GUIDELINES` and `### END_EVAL_GUIDELINES`). The surrounding prompt template, as seen in Appendix C, remains unchanged.

E.1. (iNat_butterfly)

```

'1': Compare wing patterns, colors,
markings, and body structure only
between organisms in the same life
stage.
'2': Focus on distinctive taxonomic
features rather than similarities in
pose, background, or activity.
'3': Note that similar coloration alone
doesn't indicate same species; examine
pattern arrangement and wing shape.
'4': Consider that same species will have
consistent wing pattern structure
despite variations in lighting or
angle.
'5': Different species may have convergent
color schemes but differ in pattern
details, body proportions, or wing
structure.
'6': Different life stages (caterpillar vs
adult butterfly) cannot be directly
compared to determine species identity
.
'7': When comparing different life stages,
indicate they are different species
unless specifically stated otherwise.
'8': Caterpillars of different species
often look similar; never assume a
caterpillar and butterfly are the same
species.

```

E.2. (iNat_lupine)

```

'1': Compare leaf structure, noting
palmate arrangement, leaflet count,
and size differences between species.
'2': Analyze flower color precisely -
Arctic Lupine has blue-purple flowers,
Silvery has pale blue, Arizona has
bright pink-purple/magenta flowers
that are distinctly more vibrant.
'3': Examine growth habit and overall
plant structure, as similar species
may have distinct growth forms.

```

- '4': Examine flower spike density and length - different lupine species have distinctive inflorescence patterns and heights.
- '5': Consider habitat context when visible , as different lupine species adapt to specific environments.
- '6': Look beyond superficial color similarities, as different species can share similar flower colors.
- '7': Compare leaf-to-stem ratios and branching patterns which can distinguish between lupine species.
- '8': Compare habitat context closely - Arctic prefers alpine tundra with yellow ground cover, Silvery favors mountain meadows, Arizona thrives in dry desert conditions with sandy soil.
- '9': Note flower size and individual bloom shape differences between species, beyond just color.
- '10': Examine stem characteristics - hairiness, thickness, and coloration vary significantly between lupine species.
- '11': Treat flower color differences between blue-purple and pink-purple/magenta as strong indicators of different species, especially between Arctic and Arizona lupines.
- '12': Note soil conditions - Arctic lupines grow in moist alpine environments, while Arizona lupines prefer dry, sandy desert soils.
- '13': Examine background vegetation as habitat indicators - yellow flowering ground cover often accompanies Arctic lupine in tundra environments.

E.3. (VegFru.greens)

- '1': Compare leaf shape, size, texture, color patterns, venation, and overall structure between images, noting that similar leaf features can appear across different species.
- '2': Examine flower characteristics (color , shape, arrangement) as primary identifiers when present, as these are often more species-specific than leaf features alone.
- '3': Assess whether differences are due to plant maturity, seasonal variation, or environmental conditions rather than species variation.
- '4': Prioritize reproductive structures (flowers, fruits) over vegetative features, as plants with similar

- leaves can belong to entirely different species.
- '5': Consider taxonomic features that define plant species rather than superficial similarities in appearance or growing conditions.
- '6': Pay attention to the overall growth habit, stem structure, and leaf arrangement (alternate, opposite, whorled) as additional species indicators.
- '7': Consider that plants in different growth stages may appear dissimilar while being the same species; look for consistent taxonomic markers.
- '8': Look for distinctive species-specific markers like leaf edge serration patterns, basal rosette formation, and specialized structures that differentiate similar-looking plants.
- '9': When comparing plants with only vegetative features visible, examine minute details in leaf attachment, stem characteristics, and growth patterns.

E.4. (VegFru.allium)

- '1': Compare leaf shape, thickness, color patterns, and overall structure, noting that similar-looking allium species have distinct morphological differences.
- '2': Measure relative proportions between white stem and green leaf sections, as these ratios differ consistently between species.
- '3': Distinguish between bunching onions (thin, grass-like leaves), green Chinese onions (medium thickness), and leeks (broad, flat leaves with larger white portions).
- '4': Examine bulb development - leeks have cylindrical white bases, bunching onions have minimal bulbs, Chinese onions have moderate bulb formation.
- '5': Consider overall size scale - leeks are typically larger with broader leaves than other allium species.
- '6': Note that presentation (bundled, cut, growing) doesn't change fundamental species characteristics.
- '7': Examine leaf cross-section shape (flat, tubular, or semi-circular) as this varies significantly between allium species.

'8': Compare leaf-to-stem ratio and transition zone shape, as these remain consistent within species regardless of growth stage or preparation.

E.5. (CUB_cuckoo)

'1': Compare the overall body shape, size, and proportions of the birds in both images

'2': Examine bill color precisely - Black-billed have entirely black bills, Yellow-billed have bright yellow lower mandible, Mangrove Cuckoos have grayish-yellow/buff bills with dark upper ridge

'3': Compare plumage patterns, especially noting tail feather coloration, rufous wing patches, and undertail pattern differences between species

'4': Look for eye-ring color differences - red eye-rings are distinctive for certain species

'5': Bill color differences must be carefully verified in all lighting conditions - yellow vs black vs grayish-yellow bills are definitive species indicators

'6': Pay close attention to subtle differences in head pattern, eye-ring color, and bill-to-head contrast which are key species identifiers

'7': When comparing Yellow-billed and Mangrove Cuckoos, note that similar body shape does not override the definitive bill color difference

'8': Black-billed Cuckoos have thin red eye-rings, Yellow-billed have yellow skin around eyes, Mangrove have dark/gray eye-rings - these are definitive species identifiers

'9': Verify bill color from multiple angles if possible, as shadows or lighting can obscure the true coloration

'10': When bill is partially visible or in shadow, rely on combination of eye-ring color, bill shape, and facial pattern for identification

'11': Black-billed Cuckoos have brownish upperparts, while Mangrove and Yellow-billed have grayish upperparts - this coloration difference is a key distinguishing feature

'12': When comparing similar-looking cuckoos, prioritize bill color, eye-ring color, and upperpart coloration as the three most reliable species

indicators

E.6. (CUB_oriole)

'1': Compare overall coloration patterns, noting exact distribution and intensity of colors across head, wings, breast, and tail, with special attention to contrasting color boundaries.

'2': Examine beak shape, length, and thickness as these are species-specific characteristics.

'3': Compare distinctive markings like eye rings, wing bars, throat patches, and any unique pattern that distinguishes similar-looking species.

'4': Note proportions of body parts including tail length relative to body size.

'5': Analyze color variations critically - exact hue, saturation, and distribution patterns often differentiate similar species despite general color similarities.

'6': Look for species-specific field marks even when general coloration appears similar.

'7': Compare size if reference objects are available in both images.

'8': Note habitat context if visible, as it may provide supporting evidence.

'9': When birds appear similar, prioritize diagnostic features over general appearance, as closely related species often share color schemes but differ in specific traits.

'10': Note the exact shade and intensity of similar colors, as related species may have subtly different yellow, orange, or black tones.

E.7. (CUB_vireo)

'1': 'Compare the two bird images focusing on distinctive species markers: beak shape/size, head pattern, eye color/rings, wing coloration, breast/belly color, tail features, and overall proportions.'

'2': Note subtle differences in plumage patterns, especially facial markings, throat color, and wing bars that distinguish similar-looking species.

'3': Focus on subtle differences in head pattern, eye rings, and facial markings that distinguish similar

vireo species, which often share nearly identical coloration but have species-specific facial features.

- '4': Compare birds in different poses or lighting by focusing on permanent structural features rather than variable aspects like posture or shadows.
- '5': Identify species-specific traits that remain consistent regardless of age, sex, or seasonal variation.
- '6': When birds appear highly similar, examine eye line patterns, crown coloration, and bill proportions as these often reveal definitive species differences even in similar-looking birds.
- '7': Treat similar-looking vireos (Philadelphia, Red-eyed, Warbling) as different species unless all diagnostic features match exactly, as these birds have subtle but consistent species-specific differences.
- '8': Compare the exact shape and prominence of eye rings, eyebrow stripes, and facial patterns, as these are critical for distinguishing between similar vireo species.
- '9': Note the intensity and distribution of yellow coloration on underparts, which varies consistently between similar vireo species even when overall appearance is alike.

F. Free-form Prompts

Below we show the complete generated free-form prompts for both iNat_butterfly and iNat_lupine datasets. These prompts were optimized using the same Thompson sampling and visual feedback framework as our structured YAML approach, but without structural constraints on how the agent could modify the prompt content. Additionally, we provide supplementary analysis for the lupine dataset to complement the butterfly comparison presented in the main text.

F.1. iNat_lupine: Structured vs. Freeform Comparison

The lupine classification task demonstrates even more pronounced advantages for the YAML-structured approach compared to the butterfly results presented in the main text. Our structured YAML approach achieved 0.591 accuracy compared to 0.485 accuracy for the freeform variant, representing a substantial 10.6 percentage point improvement. Additionally, the YAML format once again produced more concise prompts at 222 words versus 324 words for

freeform, achieving 1.5× greater efficiency while delivering superior performance.

The lupine domain presents particularly challenging fine-grained distinctions, as Arctic, Silvery, and Arizona lupines differ primarily in subtle petal venation patterns, leaf blade characteristics, and foliage coloration. Our structured approach proved especially effective for botanical classification due to several key advantages. The YAML format enabled botanical precision through individual guidelines that could target specific morphological features, such as "Arctic Lupine has blue-purple flowers, Silvery has pale blue, Arizona has bright pink-purple/magenta," without requiring verbose hierarchical nesting. The systematic refinement process allowed the agent to develop from generic plant comparison to species-specific diagnostic criteria, as demonstrated in our working example evolution from 0.572 to 0.644 accuracy. Furthermore, the clear feature isolation provided by numbered guidelines addressed distinct botanical aspects, enabling precise ablation studies of individual criteria.

In contrast, the freeform approach generated a verbose 5-section protocol with extensive nested subcategories. While comprehensive in scope, this structure made targeted modifications difficult because concepts like "flower color intensity" and "petal shape specifics" were distributed across multiple sections, preventing focused refinement of individual diagnostic criteria. This comparison reinforces that structured prompt evolution not only achieves better performance but produces more interpretable and maintainable classification guidelines, particularly crucial for expert domains requiring precise feature discrimination.

F.2. Freeform Prompt (iNat_butterfly):

```
# Species Identification Protocol for
  Butterfly Image Comparison

## Primary Analysis Framework
1. Focus exclusively on the butterfly's
   distinctive morphological features:
   - Wing pattern: Examine the specific
     arrangement of colors, bands, spots
     , and borders
   - Wing shape: Note the outline, edges,
     and proportional dimensions
   - Body structure: Observe thorax and
     abdomen characteristics when
     visible
   - Antennae: Note length and shape if
     visible

2. Disregard environmental factors that
   don't indicate species:
   - Background setting (plants, rocks,
     soil)
   - Lighting conditions
```

- Posture or position of the butterfly
- Image quality differences

Systematic Comparison Method

1. First Pass: Identify key distinguishing features in each butterfly
 - Document primary wing colors
 - Map pattern distribution (stripes, spots, borders)
 - Note distinctive markings or unique identifiers
2. Second Pass: Direct feature-by-feature comparison
 - Compare wing patterns: Are the specific arrangements of markings identical?
 - Compare color distribution: Do colors appear in the same pattern and proportion?
 - Compare structural elements: Are wing shapes, body proportions consistent?
3. Third Pass: Holistic evaluation
 - Would taxonomists classify these as the same species based on visible characteristics?
 - Are differences attributable to natural variation within species (gender, age, regional variation)?
 - Are differences fundamental to species classification?

Decision Rules

- Same species **REQUIRES** matching distinctive morphological features
- Superficial similarity in colors alone is **INSUFFICIENT** for same species classification
- Different wing patterns, shapes, or distinctive markings indicate different species
- When in doubt, prioritize pattern arrangement over general color impression

Remember: Similar coloration can occur across different species. The specific pattern arrangement is more taxonomically significant than general color similarity.

F.3. Freeform Prompt (iNat_Lupine):

Species Comparison Protocol for Taxonomic Differentiation

Diagnostic Feature Analysis

1. Examine flower color beyond basic hue - note specific color intensity, gradients, patterns, and markings
2. Analyze flower structure precisely - document petal shape, size proportions, and distinctive morphological features
3. Compare inflorescence architecture - observe flower density, spacing, and arrangement pattern on stems
4. Document leaf characteristics comprehensively - shape, size, margin type, venation, texture, and palmate/pinnate arrangement
5. Assess overall plant architecture and growth habit

Critical Species Differentiation Indicators

1. Flower color intensity and specific hue variations (not just "purple" but exact shade and distribution)
2. Petal shape specifics - rounded vs. pointed tips, width-to-length ratio, curvature
3. Inflorescence density and arrangement pattern (tight/loose, symmetrical/asymmetrical)
4. Leaf structure details - number of leaflets, leaflet shape, margin characteristics
5. Plant height, branching pattern, and overall growth form

Comparative Analysis Framework

1. Identify at least 5 specific diagnostic features for direct comparison
2. Document both similarities **AND** differences with equal attention to detail
3. Recognize that similar genus members (e.g., different *Lupinus* species) often have subtle but consistent distinguishing features
4. Consider that environmental factors may influence appearance but not fundamental morphological characteristics
5. Evaluate developmental stage impact on appearance (buds vs. blooms)

Decision Criteria

1. Same species: Consistent diagnostic features across both specimens despite environmental variations
2. Different species: Presence of multiple distinctive morphological differences

```

    that cannot be attributed to:
    - Environmental conditions
    - Growth/developmental stage
    - Image quality or perspective
3. When in doubt, prioritize flower
   structure and inflorescence pattern as
   primary differentiators

## Common Pitfalls to Avoid
1. Superficial color similarity without
   detailed structural analysis
2. Focusing only on obvious similarities
   while overlooking subtle diagnostic
   differences
3. Attributing all variations to
   environmental factors rather than
   taxonomic differences
4. Assuming plants with similar general
   appearance must be the same species
5. Failing to recognize that related
   species (same genus) will share family
   characteristics while maintaining
   species-specific traits

```

G. Limitations

Scale and generalizability constraints. Our evaluation focuses on fine-grained subsets with small class vocabularies (3 classes per task), which leaves the exploration of scalability to scenarios with dozens or hundreds of classes for future work. While this controlled setting enables rigorous comparison with AutoSEP and other setups, it raises questions about whether the pairwise optimization approach can maintain effectiveness as the number of classes grows.

Text-only feedback brittleness. Our experiments reveal that removing visual information leads to inconsistent performance patterns. The APO-Pair (text) variant shows non-stable behavior across batch sizes and datasets, suggesting that textual descriptions, whether generated rationales or independent image descriptions in their current state, fail to capture the subtle visual cues essential for fine-grained discrimination. This brittleness limits the method’s applicability in scenarios where visual information cannot be directly processed by the optimization agent. Future exploration of more sophisticated textual representation methods could potentially be valuable and also help us increase the number of samples fed during optimization, warranting deeper investigation.

Pair pool size exploration. Our current approach samples a pair pool offline, but the impact of pool size on optimization dynamics and final performance warrants systematic investigation. Larger pools may provide richer learning signals and better coverage of visual variations, but increase computational cost and may introduce noise. Smaller pools risk overfitting to specific image combinations and may not

capture the full spectrum of fine-grained distinctions. The optimal pool size likely depends on dataset characteristics, class similarity, and available computational budget.

Limited baseline landscape. The field of visual prompt optimization is still emerging, which constrains our ability to conduct comprehensive baseline comparisons. While existing methods like PromptAgent and CriSPO work well for text-only tasks, they weren’t designed for scenarios where the optimization agent needs to analyze visual content. In our multimodal setting, the agent must actually see the image pairs to understand classification failures—something that goes beyond what text-only optimizers can handle. Adapting these methods for visual tasks would require modifications such that the resulting systems would no longer be fully representative of the original approaches, making fair comparison difficult. This limitation reflects the early stage of multimodal prompt optimization research, where establishing robust evaluation frameworks and comprehensive baselines remains an ongoing challenge for the community.

Model evaluation. Our evaluation focuses on Qwen-2VL-7B due to organizational constraints; however, this specific model choice enables direct comparison with AutoSEP, the most relevant baseline for fine-grained visual prompt optimization. This controlled comparison isolates the impact of our pairwise contrastive approach from varying factors introduced by different model architectures. While different VLMs may exhibit varying fine-grained visual reasoning capabilities and prompt sensitivities, APO-Pair’s model-agnostic design—operating purely through inference without requiring gradient access—positions it for broad applicability. Future work should explore performance across diverse architectures like ChatGPT and Gemini Pro to further validate the method’s generalizability.