Learning to Select Visual Tools from Experience

Supplementary Material

6. More Experimental Results

6.1. Transfer to Stronger Frozen Reasoners

In the main paper (Table 2), we report single-turn transfer results where the VisTA agent, trained with QwenVL-7B as the agent model, is paired with GPT-4o as the frozen reasoning model at inference time. We further evaluate VisTA's transferability by pairing the tool-selection policy, trained only once with QwenVL-7B as the agent, with a stronger frozen reasoner, GPT-4o, at deployment time. Table 6 reports results under both single-turn and multi-turn settings.

With a single turn, VisTA achieves 88.1/75.6/52.0/55.8 on ChartQA, ChartQA-OoD, Geometry3K, and MathVerse, outperforming all training-free GPT-4o baselines. When enabling multi-turn refinement (up to three rounds), performance further improves to 88.6/79.3/53.5/57.6. These results indicate that the learned tool policy is *not tied to a specific reasoning backbone*: once trained, it transfers seamlessly to more capable VLMs without retraining. Moreover, multi-turn interaction remains beneficial even with a stronger reasoner, especially on Chart-OoD, where where queries require deeper visual understanding and robustness to imperfect tool outputs. and visual understanding is required in the absence of textual cues.

6.2. Comparison with Visual Sketchpad on Geometry3K

To further benchmark our method, we compare against Visual Sketchpad[25], a recent state-of-the-art system on Geometry3K. Visual Sketchpad enables VLMs to draw with lines, boxes, marks, which better facilitates visual perception and reasoning. For a fair comparison, we follow their evaluation protocol and test our method on the same curated subset of Geometry3K*, which is smaller than the official test set. As shown in Table7, our method still outperforms Visual Sketchpad. This result highlights the strength of our learned tool selection policy in driving effective visual reasoning.

6.3. Case Study: Multi-Turn Tool Refinement

To illustrate how multi-turn interaction improves reasoning, Fig. 7 presents a qualitative example from ChartQA. The first turn triggers a tool call to extract high-level descriptions (e.g., chart captions), but the reasoner reports low confidence (0.6), indicating insufficient information. In the second turn, the agent requests more structured tools (Raster-to-SVG and Chart-to-Table), which provide detailed geometric layouts, color information, and numerical values. With these richer signals, the reasoner produces a confident and correct answer. This example highlights how our agent leverages confidence feedback to iteratively refine tool usage rather than committing to a single static tool choice.

7. More Experimental Details

7.1. Details of Tool Pool

We construct a large, unified, and general-purpose tool pool designed to support a broad range of visual reasoning tasks. Rather than tailoring tools to individual benchmarks, we assemble a diverse ecosystem of vision and reasoning utilities that can be reused across domains, enabling the agent to learn broadly applicable tool-selection policies. The pool spans chart analysis, diagram parsing, mathematical reasoning, and low-level perception.

Chart-understanding tools: We include three categories: (1) *chart-to-table converters*: UniChart [44], DePlot [32], ChartMoE [64]; (2) *chart-to-SVG extractors*: OpenCV [9], ChartDet [66], ChartOCR [42]; and (3) *chart captioning modules*: ChartAssistant [45], ChartVLM [62], QwenVL-32B [6]. These tools provide complementary numeric, geometric, and semantic chart representations.

Geometry and symbolic reasoning tools: The pool includes *symbolic parsers*: DiagramFormalizer [71], DiagramParser [52], and *visual–symbolic solvers*: G-LLaVA [17], MultiMath [48]. While designed for geometry reasoning, these tools generalize to visually grounded mathematical tasks.

General perception tools: To broaden coverage across low-level vision tasks, we include generic visual modules such as *object detectors*: GroundingDINO [36], Mask R-CNN [23]; *depth estimators*: DepthAnything [68], MiDaS [50]; and *visual element detectors*—Harris corner detector [14], FAST [7], Canny edge detector [24], Hough line transform [69], CornerNet [30], Fast R-CNN [19].

Across categories, many tools have multiple variants differing in robustness and noise characteristics, resulting in a unified pool of 23 tools. Learning in this setting requires the agent to identify relevant tool families for each query and develop fine-grained preferences among overlapping tools, making the policy transferable across benchmarks and suitable for large-scale multimodal tool selection.

7.2. Constructing the ChartQA-OoD Test Set

Chart-based question answering poses unique challenges, requiring models to reason over numerical values, textual labels, and complex visual structures. Effective chart comprehension demands not only the identification of visual elements but also precise interpretation—such as accurately estimating bar heights in bar charts.

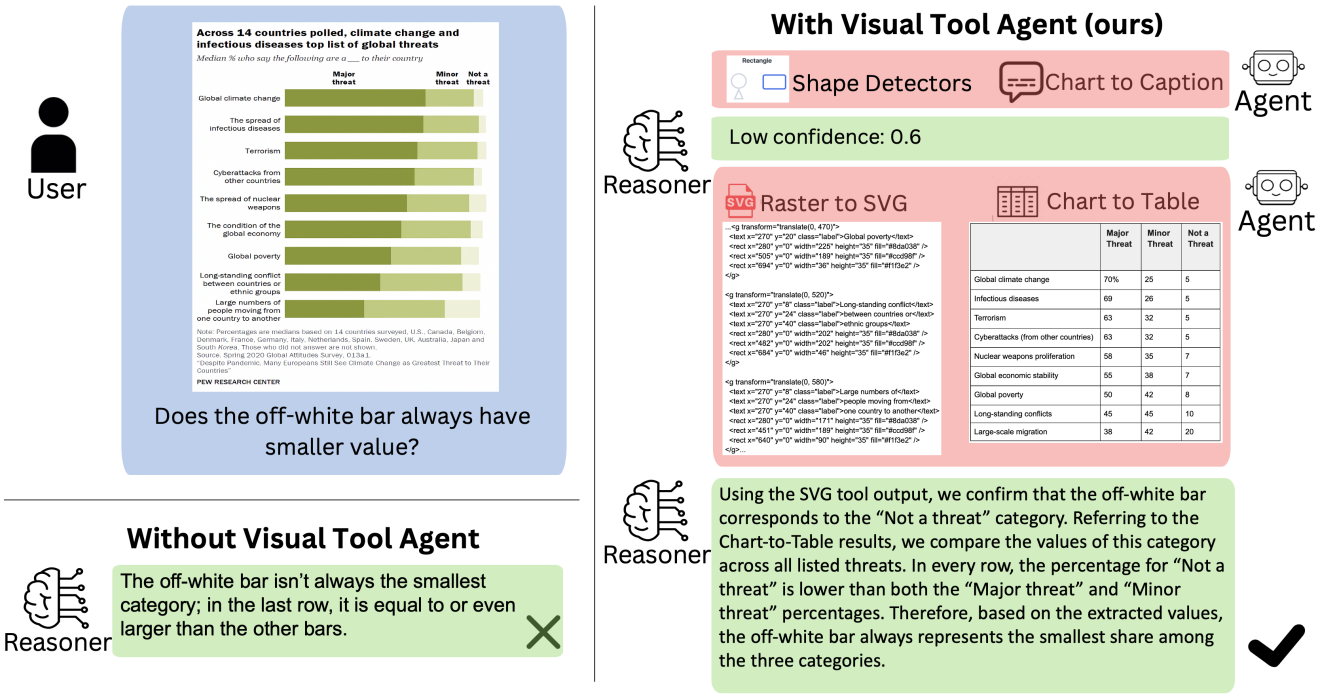

Figure 7. **Multi-turn refinement on Chart-OoD.** The agent first gathers high-level cues but receives low confidence from the reasoner, prompting a second turn that retrieves structured table + SVG information, enabling a correct answer.

Method	Agent Model	Reasoning Model	ChartQA	ChartQA(OoD)	Geometry3K	MathVerse
Training-Free	-	GPT4o	84.3	50.1	50.1	50.2
Training-Free	QwenVL 7B	GPT4o	82.3	67.1	48.7	51.4
Training-Free	GPT4o	GPT4o	84.6	73.3	49.5	53.6
Ours (Single-turn)	QwenVL 7B	GPT4o	88.1	75.6	52.0	55.8
Ours (Multi-turn)	QwenVL 7B	GPT4o	88.6	79.3	53.5	57.6

Table 6. **Transfer Results.** These results highlight VisTA's flexibility and compatibility with stronger frozen reasoning models like GPT-4o at deployment time. VisTA's tool-selection policy, trained with QwenVL-7B, transfers seamlessly to a stronger frozen reasoner (GPT-4o), yielding substantial gains across all benchmarks.

Method	Agent Model	Reasoning Model	Geometry3K*
Visual Sketchpad [25]	-	GPT4o	66.7
Ours	QwenVL 7B	GPT4o	68.8

Table 7. Comparison with Visual Sketchpad on a Geometry3K* test subset. We follow their evaluation protocol using the same curated set. *Indicates the smaller subset used by Visual Sketchpad. Our method achieves higher accuracy, demonstrating the strength of our tool selection policy.

To better evaluate the robustness and reasoning capabilities of VLMs, we construct an out-of-distribution (OoD) test set for ChartQA [43] with two targeted perturbations. We manually remove textual labels from geometric elements (e.g., bars and points) in each chart. This isolates the model's reliance on visual features by eliminating access to direct numeric answers via OCR. To further test the models' spatial reasoning robustness, we apply random geometric perturbations to the charts. With a 50% chance, each chart is either horizontally or vertically stretched to twice its original width or height. These transformations maintain semantic structure but introduce visual variability. The resulting ChartQA-OoD test set allows us to probe whether models depend heavily on textual cues and how sensitive their reasoning is to mild visual distortions. This setup provides a more rigorous benchmark for evaluating the generalization and visual understanding capabilities of VLMs.