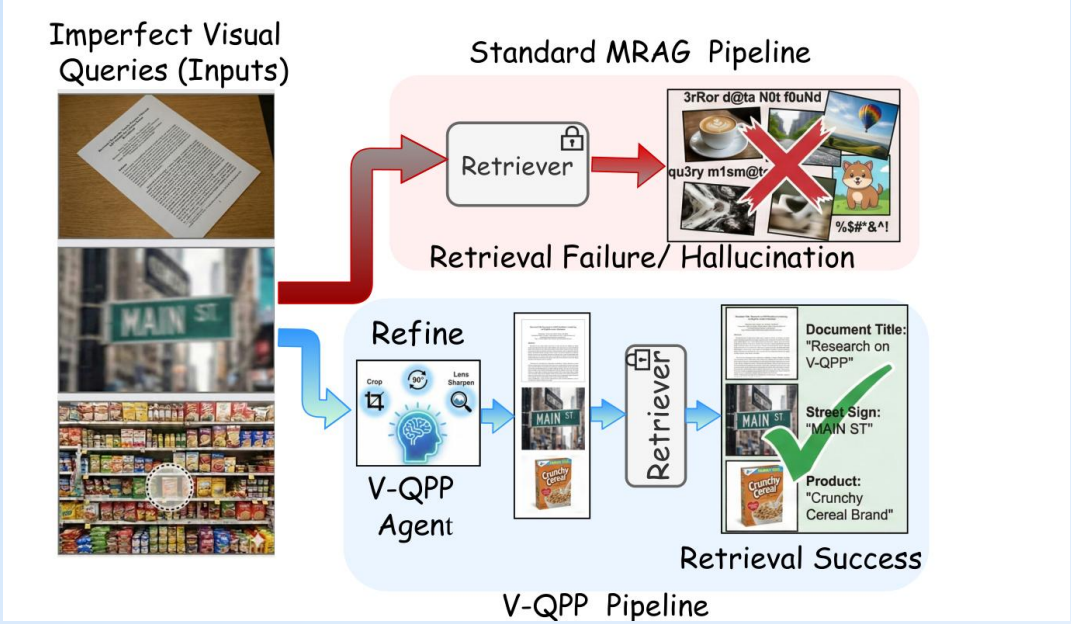
Fix Before Search

Benchmarking Agentic Visual Query Pre-processing in Multimodal Retrieval-Augmented Generation

Benchmark | **Imperfect Images**

MRAG robustness | **agentic tool use**



Presenter: Jiankun Zhang

Shenglai Zeng[2*], Jiankun Zhang[1*], Kai Guo[2 †], Xinnan Dai[2], Hui Liu[2], Jiliang Tang[1], Yi Chang[1†]

Motivation: MRAG breaks when the visual query is imperfect

Imperfect Visual Queries (Input)

Real-world queries are messy

- Geometric Distortions
  - rotated documents / flipped photos
- Quality Degradation
  - blurred signs / low-quality captures / nighttime photos
- Semantic Ambiguity
  - cluttered shelves / ambiguous targets / irrelevant watermark / occlusion

When the query image drifts away from the index representation, the retriever retrieves irrelevant knowledge, and the generator hallucinates.

Core gap: text RAG rewrites queries; MRAG rarely repairs images

Text RAG

Query rewriting, expansion, decomposition and disambiguation are standard steps before retrieval.

→

Multimodal RAG

Most pipelines treat visual inputs as static and immutable, placing the entire burden on the visual encoder.

Our question

Can an MLLM act as an active visual-query pre-processor: diagnose the imperfection, choose the right perceptual tool, and restore retrieval performance?

Shift from passive robustness to active perception

V-QPP-Bench: controlled imperfections with oracle traces

46,700 imperfect queries

10 imperfection types

5 MRAG retrieval paradigms

8 atomic operations

3 evaluation levels

Imperfections:

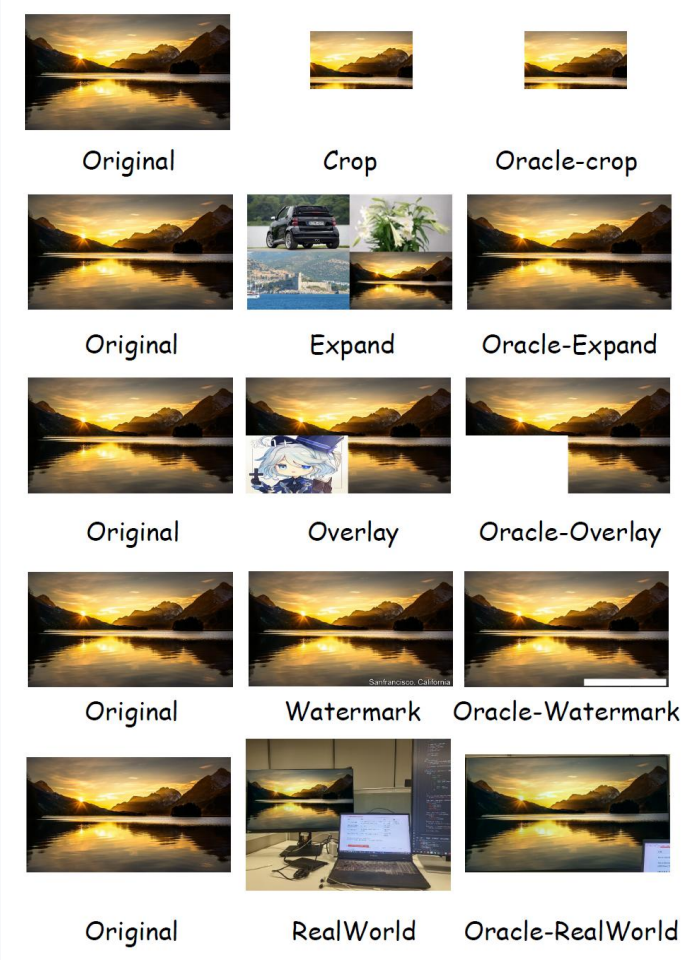
- Rotation, Flip, Brightness, Blur, Noise, Crop, Expand,
- Overlay, Watermark,
- RealWorld.

Retrieval Paradigms:

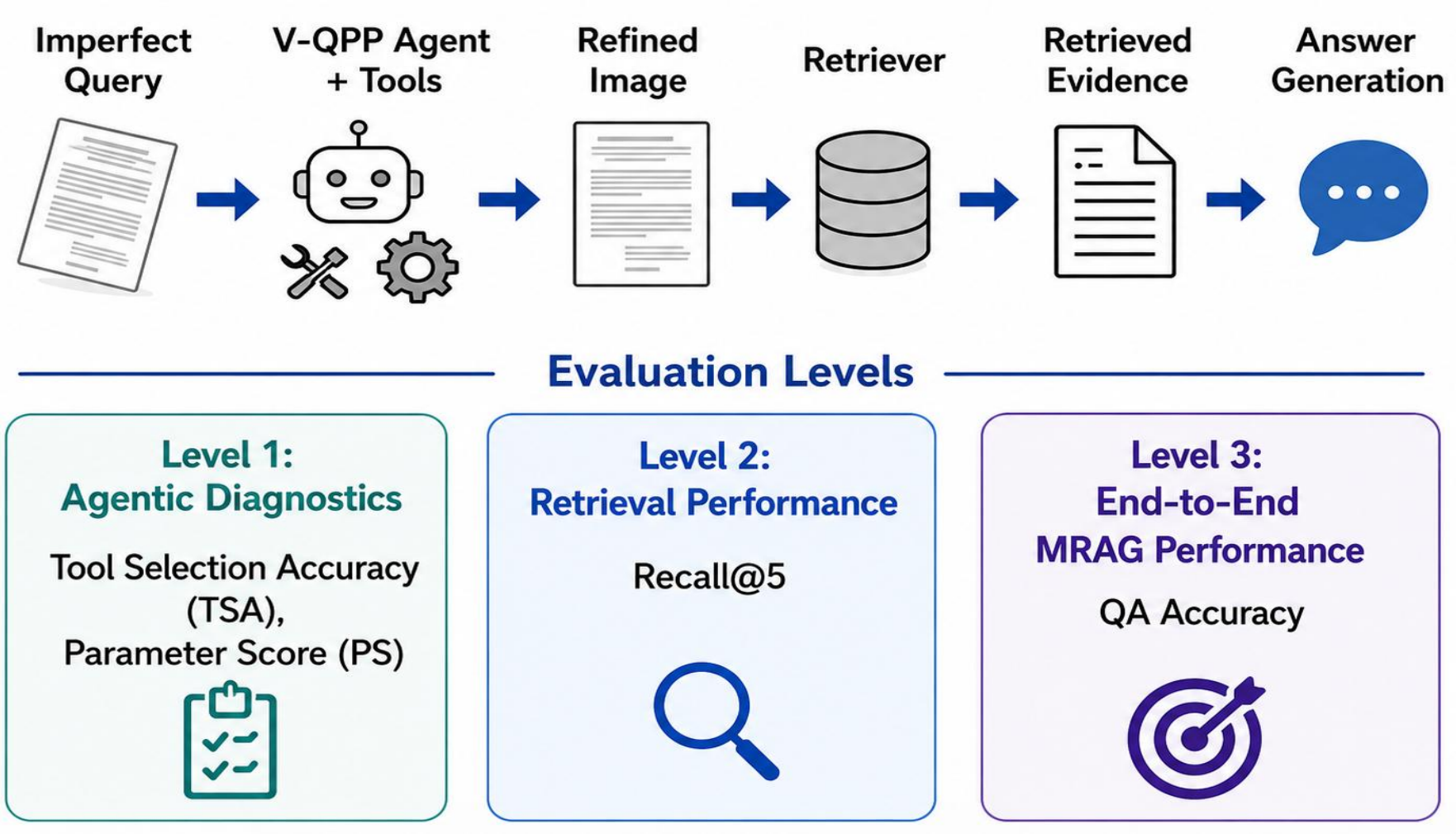
- I2T Dense Encoding
- Caption-then-Retrieve
- Image-to-Image
- Composed Retrieval
- Search API

Tool Library:

- Geometric Correction (Rotate, Flip)
- Quality Enhancement (Brightness, Deblur, Denoise)
- Semantic Refinement (Locate, Crop, Fill)



V-QPP task: diagnose, operate, retrieve



Imperfections and tools are aligned by design

Table 1. Visual corruption operations for imperfection injection.

Operation(f)	Parameters(ϕ)	Description	f^{-1} Tools
Original	None	No transformation	–
Rotation	$\theta \in \{90°, 180°, 270°\}$	Rotate by θ	$\mathcal{T}_{\texttt{rotate}}$
Flip	$\tau \in \{\text{h, v, both}\}$	Axis-aligned flip	$\mathcal{T}_{\texttt{flip}}$
Brightness	$\beta \in \{0.25, 0.5, 1.5, 1.75\}$	Scale brightness by β	$\mathcal{T}_{\texttt{lum}}$
Blur	$\sigma \in \{9, 15, 21, 27\}$	Gaussian blur (kernel σ)	$\mathcal{T}_{\texttt{deblur}}$
Noise	$\sigma_n \in \{0.05, 0.1, 0.15, 0.2\}$	Gaussian noise (σ_n)	$\mathcal{T}_{\texttt{denoise}}$
Crop	$s_c \in \{0.4, 0.5, 0.6, 0.7\}$	Crop s_c-proportional region	–
Expand	quad $\in$ {TL, TR, BL, BR}	2×2 tiling	$\mathcal{T}_{\texttt{crop}}$
Overlay	$l \in$ {TL, TR, BL, BR, C} $f \in \{0.125, 0.25, 0.5\}$	Overlay at l (scale f)	$\mathcal{T}_{\texttt{loc}}, \mathcal{T}_{\texttt{fill}}$
Watermark	font $\in \{1.0, 2.0, 3.0\}$	Text at bottom-right	$\mathcal{T}_{\texttt{loc}}, \mathcal{T}_{\texttt{fill}}$
RealWorld	template $\in \{1, 2, 3, 4\}$	Embed in screen scene	$\mathcal{T}_{\texttt{loc}}, \mathcal{T}_{\texttt{crop}}$

Table 2. Specification of atomic perceptual tools in library $\mathcal{T}$.

Tool	Parameters	Description
$\mathcal{T}_{\texttt{rot}}$	{"degrees": int}	Rotate clockwise
$\mathcal{T}_{\texttt{flip}}$	{"direction": str}	Flip (h/v/both)
$\mathcal{T}_{\texttt{lum}}$	{"factor": float}	Brightness scaling
$\mathcal{T}_{\texttt{deblur}}$	None	Remove blur
$\mathcal{T}_{\texttt{denoise}}$	None	Remove Gaussian noise
$\mathcal{T}_{\texttt{loc}}$	{"prompt": str}	Detect object, return bbox
$\mathcal{T}_{\texttt{crop}}$	{"bbox": dict}	Crop to bbox region
$\mathcal{T}_{\texttt{fill}}$	{"bbox": dict}	Fill bbox with white

Three imperfection families

Geometric | Quality | Semantic

Atomic tool library

rotate · flip · luminance · deblur · denoise · locate · crop · fill

Key challenge: successful V-QPP requires both correct tool selection and precise parameter prediction.

Evaluation: three levels, three research questions

Level 1: Agentic diagnostics

Tool Selection Accuracy (TSA) and Parameter Score (PS): did the agent choose the right operation and parameters?

Level 2: Retrieval

Recall@K: does the ground-truth evidence appear in the retrieved set after preprocessing?

Level 3: End-to-end MRAG

Substring exact-match accuracy: does the final generated answer become correct?

Research questions

- RQ1 Vulnerability: how severely do imperfect visual queries hurt retrieval and generation?
- RQ2 Capability & gap: can current MLLMs diagnose and repair these imperfections?
- RQ3 Learnability: can simple SFT teach compact models to become better V-QPP agents?

RQ1: Imperfections cause large retrieval recall drops

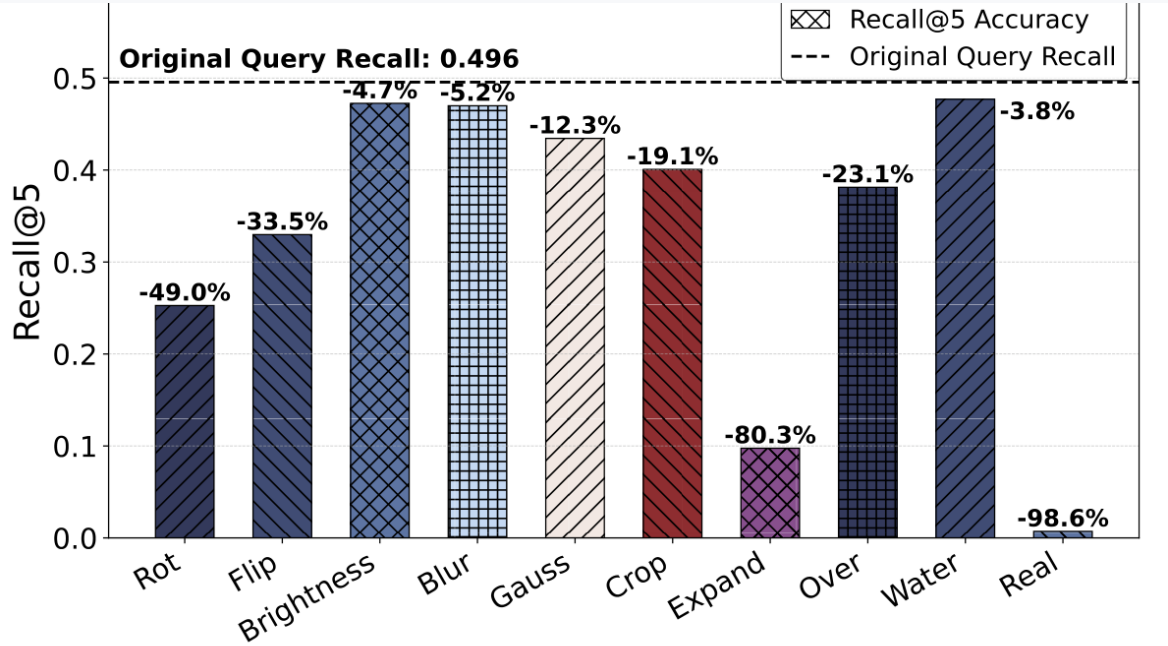
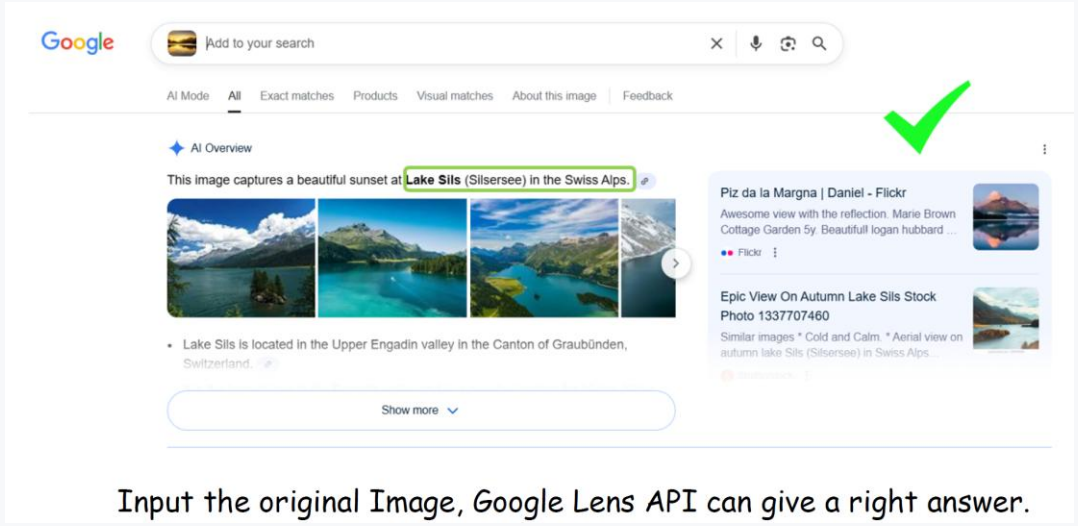
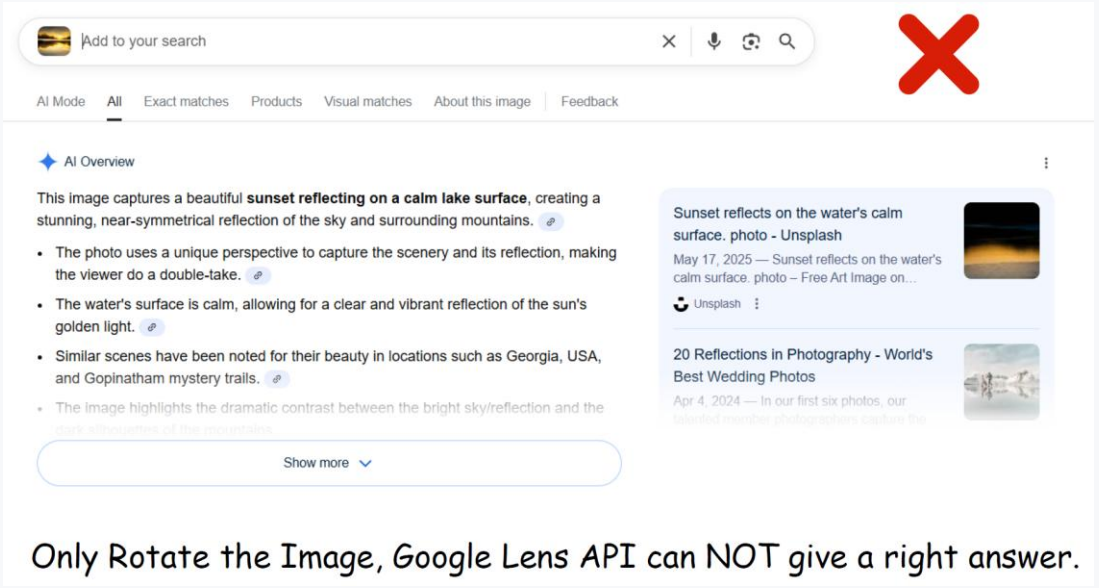


Figure 2. Retrieval performance (Recall@5) degradation in MRAG

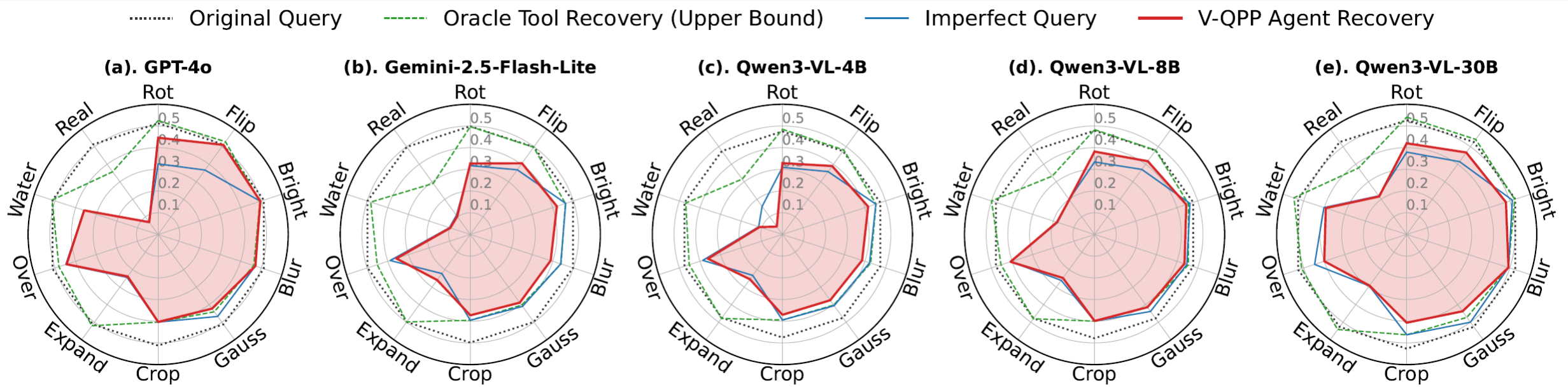
Key observation

Semantic ambiguities and geometric distortions are the most destructive. Quality degradations often matter less, but can still hurt generation quality.





RQ2: Oracle repair works, but off-the-shelf agents struggle



Restoration potential

Oracle preprocessing can recover performance for severe flaws such as RealWorld, Rotation and Flip.

Agentic bottleneck

Current MLLMs often choose the wrong tool or predict poor parameters; gains are limited and sometimes negative.

RQ3: V-QPP is learnable with simple supervised fine-tuning

Train only 1,000 imperfect queries

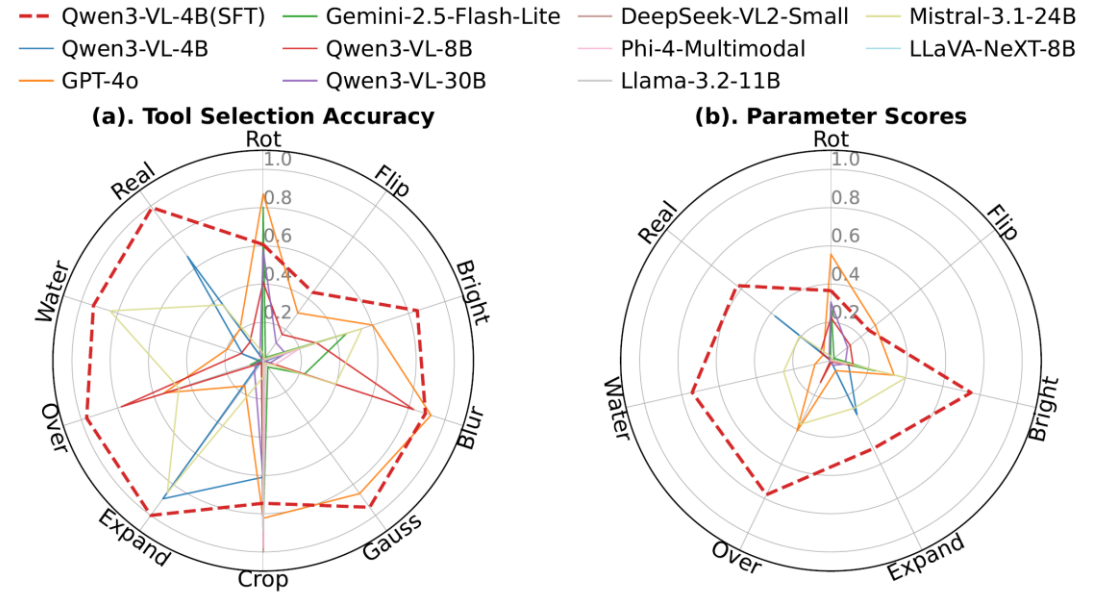


Figure 4. Tool selection accuracy and parameter scores across models. The red dot denotes Qwen3-VL-4B-Instruct after SFT, while solid lines represent off-the-shelf models.

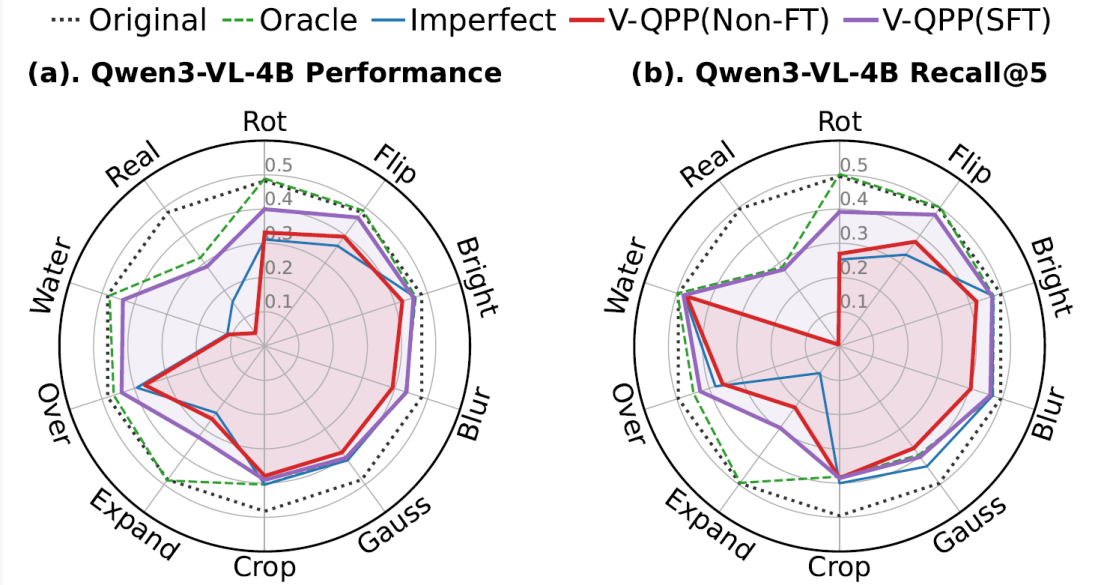


Figure 5. Performance and retrieval recall comparison. Purple line: Qwen3-VL-4B-Instruct after SFT; Red line: off-the-shelf Qwen3-VL-4B-Instruct.

After SFT, a compact 4B model matches or surpasses larger off-the-shelf models in tool usage and improves both Recall@5 and end-to-end MRAG accuracy.

TAKEAWAYS

What this paper contributes

1. New task

Defines Visual Query Pre-processing as an agentic decision-making problem before MRAG retrieval.

2. New benchmark

V-QPP-Bench covers 46.7K imperfect queries, 10 imperfection types, tool traces, and multiple MRAG paradigms.

3. New findings

Visual imperfections are highly destructive; oracle repair is promising; current agents need V-QPP-specific training.

Limitations and future directions

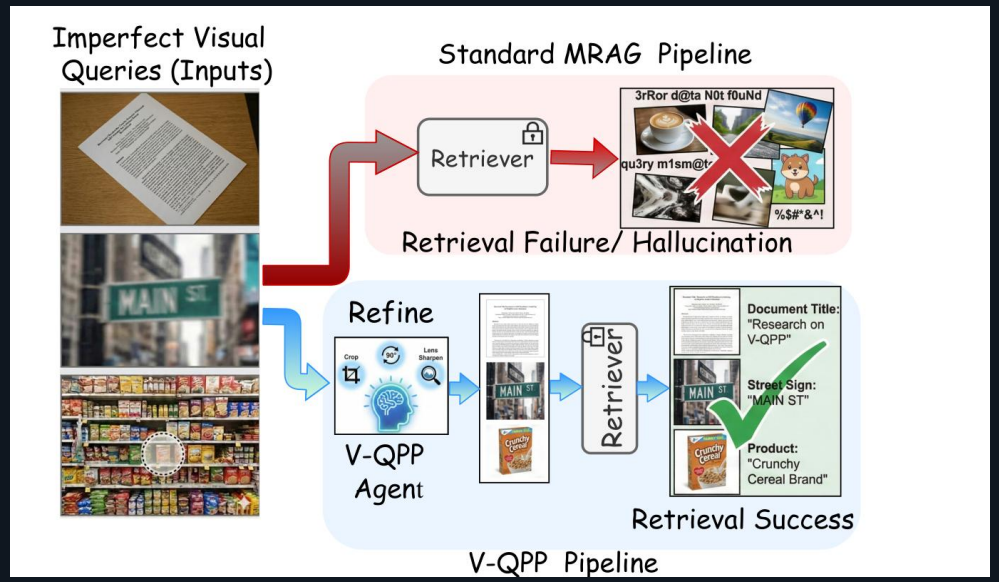
- Single-corruption setup: future work should evaluate compositional real-world defects.
- Fixed tool library: future agents may need adaptive tool discovery and multi-step planning.
- Training paradigm: SFT is a baseline; RL or outcome-aware optimization could further improve tool use.

Step(n)	Imperfect	Oracle	V-QPP(W/O SFT)	V-QPP(SFT)
1	0.323	0.384	0.310	0.368
3	0.150	0.284	0.152	0.200
5	0.056	0.180	0.062	0.097

Take-home message

Robust MRAG should not only retrieve better.
It should first see better.

Thank you!



Email: jiankun24@mails.jlu.edu.cn

V-QPP-Bench: agentic visual query preprocessing for real-world MRAG