

Supplementary Material: Benchmarking Vision Language Models as Intelligent Agents for AI Smart Glasses

Zhuohang Jiang*, Xu Yuan*, Haohao Qu, Shanru Lin, Kanglong Liu,
Wenqi Fan[†], Qing Li[†]

The Hong Kong Polytechnic University

{zhuohang.jiang, xander.yuan, haohao.qu}@connect.polyu.hk, {lllam32316, wenqifan03}@gmail.com

kl.liu@polyu.edu.hk, csqli@comp.polyu.edu.hk

We have included supplementary material to facilitate a more comprehensive understanding and in-depth analysis of the primary paper. The supplementary material is organized as follows:

- **Section A: Related Works**
- **Section B: Details of Dataset Collection**
- **Section C: Details of SUPERGLASSES**
- **Section D: Details of SUPERLENS**
- **Section E: Details of Evaluation**
- **Section F: More Experimental Results**

A. Related Works

A.1. Retrieval-Augmented VQA

Retrieval-augmented Visual Question Answering (RA-VQA) [12] extends knowledge-based VQA by coupling visual grounding with external retrieval from text corpora, KGs, or the open web (e.g., OK-VQA [20], A-OKVQA [14], WebQA [2]). Methods span knowledge-augmented transformers and RAG-enhanced LLMs to agentic/tool-use frameworks (ViperGPT [16], CRAG-MM [19]) that decide when to search, OCR, or compose multi-step programs. Recent systems adopt multi-turn, multi-hop policies that coordinate text–image retrieval and unify retrieval–generation end-to-end, learning not only to answer but to find, select, and justify evidence. Distinct from smart glasses settings, targets often occupy a very small portion of egocentric frames, making grounding and retrieval more challenging; moreover, there remains a notable lack of RA-VQA datasets specifically designed for smart glasses egocentric data.

A.2. Vision Language Model-based Agents

With the rapid advancement of large language models (LLMs) [9, 22], multimodal agents have become increas-

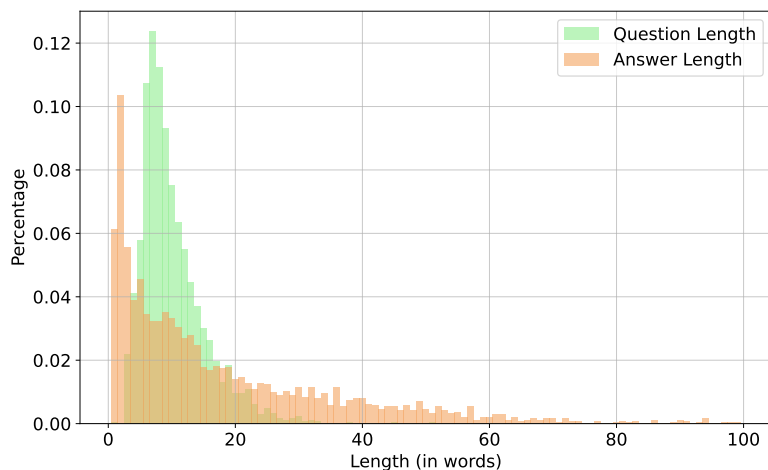
ingly capable of perceiving, reasoning, and acting across diverse modalities. Vision Language model (VLM)–based agents [17] have progressed from tool-orchestration wrappers to systems MLLM-based agents have evolved from tool orchestrators (e.g., Visual ChatGPT [21], HuggingGPT [15]) to systems that perceive–plan–act with retrieval, using prompting policies and program synthesis (ViperGPT [16]) to decompose goals and call OCR/search/vision tools. Moving beyond static images, device/UI agents operate real apps from pixels (AppAgent [23], SeeClick [4]), while embodied models (PaLM-E [5], RT-2 [24]) couple perception with action. For knowledge-heavy tasks, QA-Dragon [10] introduces a query-aware dynamic RAG, routing across text/image retrieval for multi-turn, multi-hop evidence. However, a clear gap remains: dedicated agentic stacks, purpose-built for smart glasses’ egocentric inputs, are still lacking.

B. Details of Dataset Collection

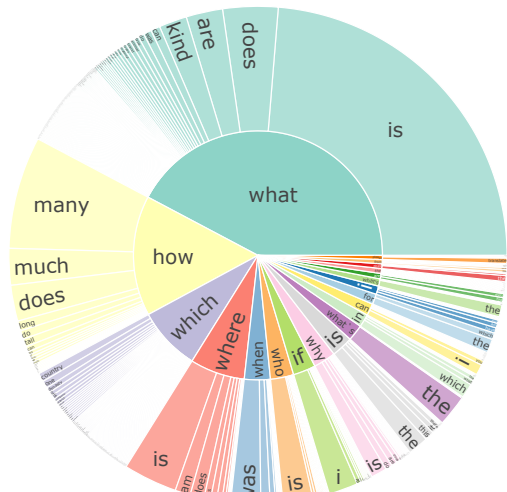
To ensure broad visual diversity, we assembled a field team of more than 20 contributors distributed across four major cities spanning three continents. Each collector was tasked with capturing high-resolution photographs in supermarkets, cafés, museums, public transit hubs, and other everyday settings, following a shared shot-list that balanced lighting conditions, camera angles, and object categories. Distinct from prior datasets, all images were captured exclusively using three mainstream smart-glasses platforms—Ray-Ban Meta, Xiaomi Smart Glasses, and RayNeo AR glasses—rather than handheld phones or DSLRs. This decision provides both device heterogeneity (different optics, sensors, and ISP pipelines) and scenario fidelity, yielding data that is intrinsically aligned with real smart-glasses usage. The collection took place across varied times of day, weather conditions, and indoor/outdoor environments to further maximise visual diversity. Before entering the annotation workflow, all raw images were

*Equal contribution.

[†]Corresponding Authors: Wenqi Fan and Qing Li



(a) Length distribution of questions and answers.



(b) Common question prefixes.

Figure A. Question/answer length distribution and question prefixes of SUPERGLASSES.

passed through a YOLO-based privacy filter¹ to automatically redact faces, license plates, and other sensitive information that may inadvertently appear in in-the-wild capture. The curated images were then uploaded to a central server and annotated via a customised Label Studio² interface with task-specific templates and scripted quality checks.

Every annotation underwent a dual-review process (collector → peer → project maintainer) to minimise label noise, standardise taxonomies, and enforce cross-annotator consistency. Category distributions were continuously monitored, and targeted sampling was employed to reduce long-tail imbalance. Collectively, these measures ensure that the resulting dataset is diverse, privacy-preserving, and faithfully representative of real-world smart-glasses VQA scenarios.

C. Details of SUPERGLASSES

C.1. Statistic Analysis

Length distribution of questions and answers. Figure Aa shows the word-length distributions for all question and answer pairs in SUPERGLASSES. Questions are typically concise and intent-driven, with a sharp peak around 7–10 words. In contrast, answers exhibit a much broader and flatter distribution, often extending beyond 50 words and occasionally reaching up to 100. This reflects the nature of smart-glasses interactions, where users issue short queries, but resolving them may require long-form reasoning, retrieval, or multi-hop tool usage. The distribution

highlights the need for agents capable of handling both succinct visual prompts and compositional, evidence-grounded answers.

Distribution of question prefixes. Figure Ab visualizes the distribution of question openers in SUPERGLASSES using a hierarchical sunburst chart. The vast majority of queries begin with “*what*”, reflecting object-centric and identity-focused information needs common in smart-glasses usage. Other frequent prefixes include “*how*”, “*which*”, and “*where*”, indicating procedural, choice-based, and location-oriented queries. Notably, the chart also reveals deeper branching structures, such as “*what is*”, “*how many*”, and “*where is*”, capturing the compositional nature of natural questions. This distribution highlights the importance of handling a wide range of query intents—from factual lookups to spatial grounding and comparative reasoning.

Distribution of common topics in questions and answers.

Figure Ba and Figure Bb present word cloud visualizations of the most frequent content words appearing in questions and answers within SUPERGLASSES. On the left, we observe that questions are heavily centered around visually grounded entities such as *building*, *flower*, *car*, *station*, and *company*, reflecting the egocentric nature of smart-glasses interactions in daily environments. The right-side word cloud shows that answers often contain semantic labels like *price*, *species*, *people*, *design*, and *water*, indicating a shift from visual recognition to external knowledge retrieval and classification. Together, the two distributions highlight

¹YOLO Model: <https://github.com/ultralytics/ultralytics>

²Label Studio: <https://labelstud.io>

the dataset’s dual emphasis on visual object grounding and multi-hop, knowledge-intensive reasoning.

C.2. Comparison with Other benchmarks

Early efforts such as WebQA [2] and InfoSeek [3] demonstrated the feasibility of multi-hop or information-seeking questions but were built on heterogeneous web crawls, lacked any notion of where the salient object sits in the frame, and provided no access to the underlying retrieval traces—making systematic analysis of failure cases or tool planning impossible. Subsequent benchmarks (MRAG-Bench [7], LIVEVQA [6], MMSearch [8], Dyn-VQA [11], and CRAG-MM [19]) expanded linguistic coverage or introduced live-update scenarios, yet they still rely on static screenshots or curated news photos, treat the object’s spatial footprint as a latent variable, and offer only “black-box” evaluation of retrieval-augmented generation without exposing the decision path.

Our SUPERGLASSES corpus addresses existing blind spots by capturing egocentric frames directly from smart-glasses users in real-world settings, precisely annotating *object coverage* to enable spatially grounded reasoning, and recording the full *search and tool-usage trajectory* to ensure that every answer is traceable to its supporting evidence. With comprehensive topical coverage across 14 distinct domains and 9 fine-grained query categories, this framework facilitates detailed result analysis, robust evaluation of retrieval and reasoning strategies, and realistic benchmarking of multimodal agents operating under the perceptual and interaction constraints of smart-glasses platforms.

C.3. Examples in SUPERGLASSES

Figures J, K, and L illustrate a four-hop example that begins with a shelf photograph of Campbell’s soup cans. The user asks “Which country is the renowned artist who painted this item from?”. Answering requires chaining visual grounding with three successive text queries:

Hop 1: Visual grounding. An image-search engine links the photo to the Wikipedia entry for *Campbell’s*, confirming the product identity.

Hop 2: Artwork retrieval. A text search for the “famous painting depicting Campbell’s” returns the article on *Campbell’s Soup Cans*, a 32-canvas pop-art series.

Hop 3: Artist identification. A follow-up query reveals that *Campbell’s Soup Cans* was painted by pop-art pioneer **Andy Warhol**.

Hop 4: Nationality verification. Consulting Warhol’s biography shows he was an **American** artist, providing the target country.

This case highlights three core properties of SUPERGLASSES: (i) *cross-modal grounding* (image \rightarrow text), (ii) *multi-document evidence aggregation* across four hops, and

(iii) *rich metadata* (device, location, difficulty, etc.) that enable fine-grained evaluation of multimodal reasoning systems.

D. Details of SUPERLENS

D.1. Demand-Adaptive Answerer

As shown in Figure 4 of the manuscript, we instruct the VLM to act as a domain router, identifying the domain of each question–image pair for domain-specific reasoning. The domain-recognition prompts are shown in Figure C, which outlines the scenarios for each predefined domain. To enable demand-adaptive retrieval, the model must also determine whether the question can be answered using its internal knowledge or requires external information. The prompt guiding this decision is illustrated in Figure D. Here, we employ chain-of-thought prompting to activate the VLM’s reasoning capabilities fully. Specifically, the VLM is instructed to identify precise object names in the image and explain their relationships to elements referenced in the question. If the VLM concludes that it has sufficient information to answer the question, it must present a comprehensive reasoning followed by the final answer. Conversely, if key information is missing, it should halt reasoning and trigger retrieval, which is subsequently handled by the Dual-Lens Knowledge Retriever.

D.2. Dual-Lens Knowledge Retriever

Given an image-question pair (v, q) , the retriever aims to extract complementary external knowledge by decomposing the retrieval process into two parallel pathways: a visual lens and a textual lens. First, a VLM-based Search Router is utilized to decide whether retrieval should rely on visual or textual signals, yielding two query sets:

$$O, Q = \text{SearchRouter}(v, q), \quad (1)$$

where $O = \{o_1, \dots, o_M\}$ are visual objects referenced in the input image and $Q = \{q_1, \dots, q_N\}$ are textual queries extracted from the original question.

For the visual branch, we employ an open-vocabulary Object Detector (Grounding DINO [13]) to localize the referred objects: $V = \{v_i = \text{ObjectDetector}(v, o_i) | o_i \in O\}$. Each detected region v_i is then used to index an image search engine, retrieving HTML pages or snippets associated with visually similar images:

$$H^{vis} = \bigcup_{i=1}^M H_i^{vis}, \quad H_i^{vis} = \text{ImageSearch}(v_i). \quad (2)$$

Because textual queries may contain multi-hop reasoning, each query q_i is further decomposed into sub-queries by a Query Decoupler, yielding an expanded set $Q' =$

Direct Answer Generation Prompt

System Prompt:

You are a visual assistant tasked with addressing the user’s query for the image based on your inherent knowledge.

General Reasoning Guidelines:

1. Generate step-by-step reasoning to address the query using evidence from the image and your knowledge... Stop reasoning once you have enough information to answer, or you find that necessary information is lacking.
2. In your reasoning, identify the exact object that the query is about by its exact name...
3. If the query involves multiple objects or relationships, dedicate one reasoning step to each object or relationship, and then summarize the result in a final step.
4. If you find that necessary information is lacking, explicitly state: “I have no knowledge about <lacking_knowledge>”

Domain Reasoning Guidelines:

...

User Prompt:

Given the <image>, please conduct step-by-step reasoning to address the query: {query}

Image metadata: The location of the image is {location}.

Output Format:

1. The exact name of the object in the image that the query is about is <specific_object_name>.
2. Then, ...
3. Therefore, the answer is ...

Output Summary in JSON format: {“reasoning”: <summary_reasoning_string>, “answer”: <answer>}

Figure D. The prompt used for direct answer generation.

E. Details of Evaluation

E.1. LLM-as-Judge

As shown in Figure E, we employ a structured evaluator prompt to assess answer correctness via an LLM-based evaluator, i.e., Qwen2.5-32B. The evaluator is instructed to act as an expert QA judge and is given clear reasoning guidelines. It determines whether a predicted answer is accurate by comparing it against the ground truth, allowing for surface-level variation (e.g., paraphrasing) as long as the semantic content is preserved. The prompt enforces strict criteria: missing key details or including incorrect information results in a negative judgment. The output is a JSON object with a single Boolean field, *accuracy*, which enables consistent downstream aggregation and scoring.

E.2. Settings of Direct Answering






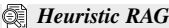
In the *Direct Answering* setting, we adopt a concise VQA prompting scheme that forgoes chain-of-thought reasoning and external retrieval. The model is guided only by a lightweight system prompt that constrains its behavior (e.g., “You are a helpful assistant that answers questions based on the provided image.”). During inference, we directly concatenate the encoded image and user query with a brief user prompt (“Please give the answer within 1–2

sentences. Answer:”), which encourages the model to produce a short, image-grounded response relying solely on its vision–language understanding and inherent knowledge. To reduce stochasticity in next-token prediction, we set the sampling temperature of all VLMs to 0.

E.3. Settings of Heuristic RAG

In the *Heuristic RAG* setting, we employ a straightforward retrieval pipeline that relies solely on external search engines, without incorporating any learned retrieval components. For each user query, we directly submit the textual query to Google Search and collect the top-ranked web snippets as candidate evidence. In parallel, the input image is processed using Google Lens, which returns visually similar webpages along with their associated metadata. All retrieved webpages are then passed through the Webpage Reader used in our SUPERLENS system to extract clean textual content. The resulting snippets are concatenated into a lightweight context buffer and provided to the model without any reranking, filtering, or structured reasoning. For *Text RAG*, we use only the textual snippets retrieved from the original query. For *Image RAG*, we use only the image-relevant results returned by Google Lens. Overall, this heuristic design approximates a naïve “search-then-answer” pipeline, allowing us to isolate how much performance can

Table A. Detailed scores on top-10 domains of SUPERGLASSES.

Model	Plant	Public Service	Food	Shopping	Translation	Transport	Culture	Navigation	Animal	Education
										
LLaMA-3.2-11B	22.49	24.58	18.82	28.27	25.11	17.24	25.13	30.46	18.35	30.00
LLaMA-3.2-90B	34.32	33.00	26.20	34.60	34.04	26.11	33.51	35.06	25.32	37.50
										
Qwen2.5-VL-3B	26.04	28.96	20.30	29.96	25.11	20.20	25.13	33.91	10.76	40.00
Qwen2.5-VL-7B	31.07	34.01	22.51	35.44	39.57	25.12	26.70	48.28	25.95	45.00
Qwen2.5-VL-32B	37.87	36.36	29.89	40.08	43.83	26.11	36.65	44.83	28.48	50.00
Qwen2.5-VL-72B	37.87	39.73	28.78	42.62	43.40	31.03	34.03	48.85	31.65	46.25
										
MiMo-VL-7B	26.63	25.93	20.30	28.69	21.70	15.76	27.23	32.76	18.35	37.50
										
Phi-3-Vision-4B	18.34	19.53	15.13	18.57	16.60	11.33	16.23	21.84	11.39	27.50
InternVL3-8B	24.26	30.98	19.19	32.49	28.51	15.27	26.18	36.78	22.15	40.00
GLM-4.1V-9B	22.78	23.57	21.03	27.43	22.98	19.70	27.23	27.01	18.35	35.00
LLaVA-v1.5-7B	11.54	9.76	11.44	12.66	7.23	6.90	12.57	13.22	6.96	12.50
LLaVA-v1.5-13B	13.61	10.44	12.55	16.03	5.11	9.36	14.66	12.07	9.49	15.00
LLaVA-Onevision-0.5B	10.65	16.50	12.92	18.14	7.66	12.81	14.14	16.67	13.29	25.00
LLaVA-Onevision-7B	19.82	21.21	15.87	24.47	17.45	16.26	20.42	31.61	14.56	36.25
DeepSeek-VL2-3B	17.46	20.54	17.71	21.94	16.60	18.72	19.37	29.31	15.19	35.00
DeepSeek-VL2-16B	24.26	23.23	21.03	28.27	21.70	18.72	24.08	32.76	17.09	33.75
DeepSeek-VL2-27B	23.08	28.96	21.77	31.65	25.53	17.73	26.18	33.91	18.99	40.00
										
GPT-4o	46.75	43.10	40.22	44.30	42.55	33.99	38.74	44.83	39.87	53.75
Claude 4 Sonnet	31.95	37.37	38.38	42.19	41.28	29.56	36.13	43.10	27.22	53.75
Gemini 2.5 Pro	40.83	46.13	42.80	46.41	38.30	41.38	39.79	48.85	36.08	53.75
										
LLaMA-3.2-11B	14.50	15.49	14.02	24.89	15.32	14.29	19.90	20.69	10.76	36.25
LLaMA-3.2-11B	11.54	12.12	12.92	14.77	12.77	12.32	16.23	14.94	8.23	20.00
LLaMA-3.2-11B	12.72	10.77	14.39	16.46	14.04	13.30	19.90	19.54	9.49	26.25
Qwen2.5-VL-7B	15.09	19.87	19.19	21.52	20.00	13.30	20.94	26.44	12.66	41.25
Qwen2.5-VL-7B	16.57	15.15	15.50	17.72	17.87	13.30	17.80	19.54	10.13	26.25
Qwen2.5-VL-7B	16.57	17.17	14.39	21.10	17.87	12.32	19.37	20.69	12.03	26.25
SUPERLENS[†] (Ours)	33.43	36.03	32.10	48.52	33.62	33.00	40.31	48.28	36.08	41.25
	↑10.94	↑11.45	↑13.28	↑20.25	↑8.51	↑15.76	↑15.18	↑17.82	↑17.73	↑11.25
SUPERLENS[‡] (Ours)	39.64	47.47	34.32	47.68	51.91	38.42	38.74	56.90	34.81	56.25
	↑8.57	↑13.46	↑11.81	↑12.24	↑12.34	↑13.30	↑12.04	↑8.62	↑8.86	↑11.25

be attributed to simple external retrieval signals versus more advanced retrieval-augmented generation mechanisms.

F. More Experimental Results

In this section, we present the detailed scores of 26 leading VLMs alongside our SUPERLENS across the top 10 image domains (Table A) and 8 query categories (Table B). The results show that our method consistently achieves substantial improvements across all evaluated scenarios. Notably, SUPERLENS[‡] surpasses the performance of large-scale models such as Gemini 2.5 Pro and GPT-4o, both estimated to contain more than 400B parameters, on multiple domains and query categories.

Table C presents detailed results across the Dynamism and Reasoning Hops dimensions. We observe that both SUPERLENS[†] and SUPERLENS[‡] achieve substantially larger improvements in the Fast-Changing setting than in the Static or Slow-Changing settings, highlighting the effectiveness of our approach for queries whose answers shift rapidly over time. Additionally, while LLaMA-3.2-11B with direct answering performs poorly on 3-hop and 4-hop questions, its performance improves markedly when augmented with our method (SUPERLENS[†]), showing gains of more than 17% and 21%, respectively. This pattern indicates that SUPERLENS is highly capable of tackling complex, multi-step questions even when the backbone model is relatively weak,

Evaluator Prompt

System Prompt:

You are an expert evaluator of question-answering systems.

User Prompt:

General Reasoning Guidelines: “Your task is to determine if a prediction correctly answers a question based on the ground truth.”

Rules:

1. The prediction is correct if it captures all the key information from the ground truth.
2. The prediction is correct even if phrased differently as long as the meaning is the same.
3. The prediction is incorrect if it contains incorrect information or is missing essential details. “Output a JSON object with a single field ‘accuracy’ whose value is true or false.”

Question: {query}, *Ground Truth:* {answer}, *Prediction:* {prediction}

Figure E. The prompt used for answer evaluation.

a capability largely attributable to the Query Decoupler and the fine-grained RAG pipeline.

Furthermore, as table D shows, we selected the LLM-as-judge metric for its consistency, scalability, and ability to capture semantic meaning—advantages that conventional measures such as lexical or embedding similarity often lack. Meanwhile, we extend our metrics to include LLMs from diverse families, hybrid LLM-based evaluation, and human assessments to mitigate potential bias. Also, We have incorporated the suggested multilingual analysis, evaluating performance on English (en), Chinese (zh), French (fr), and Japanese (jp) in our main experiments, as shown in Table E

F.1. More Findings

Humans and vision language models often diverge in their perception of question difficulty and the necessity of external retrieval. Tasks that appear easy for humans may challenge models without additional context, while some seemingly complex questions for humans are handled well by models through memorized patterns. Notably, stronger models like GPT-4o and Gemini 2.5 Pro rely less on retrieval, maintaining high accuracy even without external input, suggesting greater internal world knowledge. In contrast, weaker models benefit more from retrieval but are also more sensitive to irrelevant results.

F.2. Case Study

To better understand how our system behaves across different multimodal reasoning scenarios, we conduct a targeted case study that examines both its successes and failures in real examples. Specifically, we present two success cases and two failure cases to analyze the end-to-end decision process, from tool invocation to query decoupling to evidence aggregation. The success cases highlight situations where the model accurately identifies the appropriate retrieval modality (image search vs. text search) and con-







structs well-formed queries that lead to reliable answers. In contrast, the failure cases reveal two major failure modes: incorrect selection of the search tool and poorly structured query decoupling. Together, these examples provide a fine-grained, qualitative view of how the system makes decisions, what it does well, and where the current limitations lie.

F.2.1. Success Case

To illustrate how our system behaves under ideal conditions, we present two representative success cases, each showcasing a different but appropriate tool invocation strategy. In the first case (Figure F), the model is asked to identify the vehicle model from an image. The system first performs precise visual grounding by extracting the relevant region through object detection, and then issues an image-based search query tailored to the cropped vehicle. This example illustrates that the model not only selects the correct modality (image search instead of text search) but also constructs a semantically faithful and discriminative query, enabling the retriever to return the correct result for the “Honda Freed Hybrid”.

In contrast, the second case (Figure G) demonstrates a query whose answer is not visually observable, i.e., the number of founders of the Sushiro restaurant chain. Here, the model successfully recognizes that the image provides insufficient information and therefore switches to text search. It generates a concise and meaningful search query (“Sushiro founders Hong Kong”) aligned with the question intent, retrieves high-quality textual evidence, and produces the correct answer (“two founders”). Together, these two examples show that when the system correctly selects the appropriate search tool, forms a well-structured query, and extracts modality-specific evidence, it is capable of producing robust and accurate answers across heterogeneous question types.

Table B. Detailed scores on query categories of SUPERGLASSES.

Model	Aggregation	Comparison	Factual Knowledge	Multi-hop	Reasoning	Simple Recognition	Spatial Reasoning	Temporal Understanding
 <i>Meta-Rayben Smart Glasses</i>								
LLaMA-3.2-11B	19.42	21.23	18.09	16.79	32.38	24.53	33.94	21.54
LLaMA-3.2-90B	26.62	32.88	26.50	23.80	38.95	34.08	38.53	32.31
 <i>RayNeo Smart Glasses</i>								
Qwen2.5-VL-3B	22.30	28.08	17.52	17.52	31.03	30.43	31.19	24.62
Qwen2.5-VL-7B	30.22	34.93	21.94	23.94	41.48	39.33	34.86	30.77
Qwen2.5-VL-32B	29.50	39.73	26.21	27.30	48.57	42.13	39.45	38.46
Qwen2.5-VL-72B	37.41	42.47	28.21	28.76	50.25	42.32	37.61	35.38
 <i>XiaoMi Smart Glasses</i>								
MiMo-VL-7B	25.18	27.40	16.67	15.77	29.51	29.12	35.78	27.69
 <i>Open-sourced VLMs</i>								
Phi-3-Vision-4B	12.23	27.40	11.68	11.09	24.79	18.63	24.77	21.54
InternVL3-8B	24.46	32.19	16.81	19.85	36.76	30.99	34.86	30.77
GLM-4.1V-9B	22.30	28.77	16.52	16.35	31.87	27.72	24.77	21.54
LLaVA-v1.5-7B	10.79	17.12	6.41	7.88	13.15	11.89	20.18	12.31
LLaVA-v1.5-13B	13.67	20.55	7.98	8.03	13.49	12.17	24.77	20.00
LLaVA-Onevision-0.5B	13.67	17.81	7.98	8.18	16.69	16.57	18.35	18.46
LLaVA-Onevision-7B	20.86	30.82	13.53	14.45	26.14	23.69	34.86	30.77
DeepSeek-VL2-3B	17.27	28.08	11.54	11.39	26.64	23.69	26.61	24.62
DeepSeek-VL2-16B	23.74	30.14	14.67	15.62	29.51	29.96	27.52	24.62
DeepSeek-VL2-27B	29.50	34.93	22.22	19.85	34.57	32.96	37.61	32.31
 <i>Proprietary VLMs</i>								
GPT-4o	34.53	41.78	37.89	34.89	50.08	44.48	52.29	38.46
Claude 4 Sonnet	35.25	38.36	29.06	29.49	49.75	41.01	41.28	43.08
Gemini 2.5 Pro	41.73	51.37	38.18	38.39	50.25	45.13	43.12	36.92
 <i>Heuristic RAG</i>								
LLaMA-3.2-11B	11.51	17.81	12.68	11.97	23.27	19.76	22.02	26.15
LLaMA-3.2-11B	12.95	21.92	12.25	9.49	16.53	13.01	14.68	21.54
LLaMA-3.2-11B	15.11	25.34	12.68	9.64	18.21	14.61	14.68	26.15
Qwen2.5-VL-7B	16.55	20.55	13.82	12.12	27.15	22.94	16.51	24.62
Qwen2.5-VL-7B	12.23	19.18	13.82	10.07	20.57	17.04	15.60	23.08
Qwen2.5-VL-7B	12.23	20.55	13.68	9.78	20.57	18.63	19.27	23.08
SUPERLENS[†] (Ours)	27.34 ↑7.92	43.84 ↑22.61	37.61 ↑19.52	30.80 ↑14.01	42.16 ↑9.78	36.99 ↑12.46	43.12 ↑9.18	32.31 ↑10.77
SUPERLENS[‡] (Ours)	39.57 ↑9.35	45.89 ↑10.96	36.47 ↑14.53	34.74 ↑10.80	52.28 ↑10.80	49.34 ↑10.01	44.95 ↑10.09	44.62 ↑13.85

F.2.2. Failure Case

In addition to the success examples, we further present two failure cases (Figures H and I) to highlight the system’s failure modes and the underlying causes.

The first case (Figure H) illustrates a mis-selection of the search tool. The question asks about a red symbol visible on a glass window. Instead of performing image-based search, which is necessary because the symbol is visually specific, hard to describe verbally, and not uniquely identifiable by text, the system mistakenly triggers text search with the query “red sign meaning in Canada.” This textual query is overly broad and semantically disconnected from the object in the image, leading the retriever to return ir-

relevant knowledge about the “Canadian Red Ensign” flag. This failure demonstrates that when the model misjudges the modality of the problem and chooses the wrong retrieval channel, even a formally well-constructed text query cannot compensate for the mismatch between modality and information need.

The second case (Figure I) showcases a different type of failure: incorrect search query generation. The question asks, “What other books has the author of this book written?” The image clearly shows only the book cover and bookshelf, meaning the system should first infer the author’s name visually and then construct a clean, author-centric search query. However, the model generates a noisy

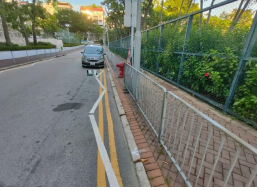

 <p>Query: What is the model of this vehicle?</p> <p>Reasoning:</p> <ol style="list-style-type: none"> The exact name of the object in the image that the query "What is the model of this vehicle?" is about is a car. The car in the image appears to be a Toyota model, as indicated by the logo on the front. The specific model cannot be determined from the image alone without more details such as the body style or additional features. Therefore, the answer is: {"reasoning": "The model of the vehicle in the image is a Toyota, but the specific model cannot be determined from the image alone."} <p>Decision: Need RAG</p> <p>Tool Decision: Need Image Search, No need Text Search</p>	<p>Object Detection: Car</p>  <p>Search Result:</p> <pre>"html_search": { "position": 1, "title": "🎉 NEW ARRIVAL! 🎉 Discover Your Dream Ride ...", "link": "https://www.facebook.com/kromecars/posts/-NEW-ARRIVAL-discover-your-dream-ride-in-singapore-honda-freed-hybrid-15a-g-7-se/721948503911114/", "source": "Facebook", "thumbnail": "https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcTW-rzltcBS2bio41Lg2Pec-HN2H6J0GNGDf7eTbeJ7gAV_UDoA", "thumbnail_width": 259, "thumbnail_height": 194, "image": "https://lookaside.fbsbx.com/lookaside/crawler/media/?media_id=721947653911199", "image_width": 1600, "image_height": 1200 },...</pre>	<p>RAG Result:</p> <p>Evidence 1 with relevance score 0.88:</p> <p>July 10 · Discover Your Dream Ride in Singapore! Honda Freed Hybrid 1.5A G 7-Seater Honda Sensing Registered: 21/08/2020 Road Tax: \$682/year COE: \$35,001 OMV: \$29,003 ARF: \$22,605 No. of Owners: 1 Key Features: ...</p> <p>Evidence 2 with relevance score 0.88:</p> <p>For sale Tokyo Japan 2009 Honda Freed</p> <hr/> <p>Mick Lay Mick Lay 11.2K subscribers Subscribe Subscribed</p> <p>Reasoning: The evidence mentions the Honda Freed Hybrid model, which is described as a 7-seater with features like a 7-speed DCT transmission and Honda Sensing technology.</p> <p>Answer: The model of the vehicle is the Honda Freed Hybrid.</p> <p>Ground Truth: Honda Freed.</p>
---	---	---

Figure F. Success case of using image search.


 <p>Query: How many founders founded Sushiro?</p> <p>Reasoning:</p> <ol style="list-style-type: none"> The exact name of the object in the image that the query "How many founders founded Sushiro?" is about is Sushiro. Sushiro is a Japanese restaurant chain. The founders of Sushiro are not visible in the image. I cannot determine the number of founders of Sushiro based on the image. <p>Therefore, the answer is: {"reasoning": "I cannot determine the number of founders of Sushiro based on the image."}</p> <p>Decision: Need RAG</p> <p>Tool Decision: No need Image Search, Need Text Search</p>	<p>Search Query: Sushiro founders Hong Kong</p> <p>Search Result:</p> <pre>"html_search": { "position": 1, "title": "Sushiro", "link": "https://en.wikipedia.org/wiki/Sushiro", "thumbnail": "https://serpapi.com/searches/68fcd0b56838633e892da973/...", "favicon": "https://serpapi.com/searches/...", "snippet": "Founder, Yoshio Shimizu Yutaka Shimizu. Headquarters, 〒564-0063. ...", "snippet_highlighted_words": ["Yoshio Shimizu Yutaka Shimizu"], "source": "Wikipedia" }, { "position": 4, "title": "FOOD & LIFE COMPANIES LTD. / Sushiro", "link": "https://hk.linkedin.com/company/food-and-life-companies-ltd/", "favicon": "https://serpapi.com/searches/68fcd0b56838633e892da973/...", "snippet": "Masahiro Yamamoto. CEO, FOOD & LIFE COMPANIES LTD.; Eishinta Uinsawati...", "snippet_highlighted_words": ["Masahiro Yamamoto. CEO, FOOD & LIFE COMPANIES LTD"], "source": "LinkedIn - FOOD & LIFE COMPANIES LTD. / Sushiro" },</pre>	<p>RAG Result:</p> <p>Evidence 1 with relevance score 0.94:</p> <p>From Wikipedia, the free encyclopedia Akindo Sushiro Co., Ltd. Sushiro headquarters Company typeKabushiki gaisha IndustryRetail (Japan) Food service (other countries) Founded October 23, 1984; 40 years ago Founder Yoshio Shimizu Yutaka Shimizu Headquarters: 564-0063 ...</p> <p>Evidence 2 with relevance score 0.87:</p> <p>QTS Scheme restaurant SUSHIRO, founded over 30 years ago, is the leading kaiten sushi chain in Japan, according to the Fuji Keizai Food service Industry Marketing Data Book 2012 – 2022. Choose from an extensive menu with more than 100 delightful options, including sushi and side dishes, all offered at reasonable prices.</p> <p>Reasoning: The evidence states that Sushiro was founded by Yoshio Shimizu and Yutaka Shimizu, indicating that two people founded the company.</p> <p>Answer: Two people founded Sushiro.</p> <p>Ground Truth: 2.</p>
--	---	--

Figure G. Success case of using text search.


 <p>Query: What does this red sign mean?</p> <p>Reasoning:</p> <ol style="list-style-type: none"> The exact name of the object in the image that the query "What does this red sign mean?" is about is a red line on a glass window. The red line on the glass window is likely a safety feature, such as a grab bar or a design element. In many high-rise buildings, red lines on glass windows are often used as safety features to help people hold onto the glass in case of an emergency. The red line could also be a design element, possibly part of the building's branding or aesthetic. <p>Therefore, the answer is: {"reasoning": "The red line on the glass window is likely a safety feature, such as a grab bar or a design element."}</p> <p>Decision: Need RAG</p> <p>Tool Decision: No need Image Search, Need Text Search</p>	<p>Search Query: red sign meaning in Canada</p> <p>Search Result:</p> <pre>"html_search": { "position": 1, "title": "Canadian Red Ensign", "link": "https://en.wikipedia.org/wiki/Canadian_Red_Ensign", "favicon": "https://serpapi.com/searches/690da69bfbbeb5d069a23281/images/dcf4c1e0b2efa285e4bccc2e36e1be59fa49c64aeb9a8679b097c1000628e274c.png", "snippet": "During the Second World War, the ensign saw use as symbol that represented Canada's armed forces. In 1945, an order in council named the Canadian Red Ensign a '...', "snippet_highlighted_words": ["represented Canada's armed forces"], "source": "Wikipedia" }, { "position": 8, "title": "Bellow Ontarians, you see this sign, what do you do if ...", "link": "https://www.reddit.com/r/ontario/comments/zufopi/fellow_ontarians_you_see_this_sign_what_do_you_do/", "favicon": "https://serpapi.com/searches/690da69bfbbeb5d069a23281/images/dcf4c1e0b2efa285e4bccc2e36e1be59fa49c64aeb9a8679b097c1000628e274c.png", "snippet": "The sign is instructing you where to come" }</pre>	<p>RAG Result:</p> <p>Evidence 1 with relevance score 0.85:</p> <p>From Wikipedia, the free encyclopedia \n Canadian Red Ensign.svg \n Red Ensign used from 1957 to 1965 \n UseCivil ensign \n Proportion 1:2 \n Adopted 1957 (earlier versions in 1892 and 1922) \n Relinquished 1965 \n Design A red field, featuring the Royal Union \n Flag in the canton "Canton (flag)", defaced "Defacement (flag)" with the shield portion of the coat of arms of Canada.</p> <p>Evidence 2 with relevance score 0.82:</p> <p>Design A red field, featuring the Royal Union Flag in the canton "Canton (flag)", defaced "Defacement (flag)" with the shield portion of the coat of arms of Canada. The Canadian Red Ensign served as a nautical flag and civil ensign for Canada from 1892 to 1965, and later as the ..., emblazoned with the shield of the coat of arms of Canada.</p> <p>Reasoning: The Canadian Red Ensign is a nautical flag and civil ensign for Canada, featuring the Royal Union Flag in the canton with the shield of the coat of arms of Canada, and it was used from 1892 to 1965.</p> <p>Answer: The red sign is the Canadian Red Ensign, a flag used as a nautical and civil ensign for Canada from 1892 to 1965.</p> <p>Ground Truth: This red emblem is the Toronto CN Tower's logo.</p>
---	--	--

Figure H. Failure case caused by incorrect search tool selection.

Table C. Detailed scores on dynamism and hop dimensions of SuperGlasses.







Model	Serach Type	Dynamism			Reasoning Hops			
		Static	Slow-Changing	Fast-Changing	1-hop	2-hop	3-hop	4-hop
 <i>Meta-Rayben Smart Glasses</i>								
LLaMA-3.2-11B	Direct Answer	25.21	13.66	13.92	26.35	23.53	16.90	4.35
LLaMA-3.2-90B	Direct Answer	32.80	22.36	26.29	34.55	29.41	24.80	17.39
 <i>RayNeo Smart Glasses</i>								
Qwen2.5-VL-3B	Direct Answer	42.59	25.77	18.81	29.02	21.85	17.69	13.04
Qwen2.5-VL-7B	Direct Answer	34.64	23.60	21.13	36.92	27.73	23.70	17.39
Qwen2.5-VL-32B	Direct Answer	37.93	26.09	30.41	40.50	33.61	27.33	21.74
Qwen2.5-VL-72B	Direct Answer	38.85	27.95	32.99	41.29	36.13	29.07	21.74
 <i>XiaoMi Smart Glasses</i>								
MiMo-VL-7B	Direct Answer	26.08	16.77	20.62	28.96	15.97	16.90	13.04
 <i>Open-sourced VLMs</i>								
Phi-3-Vision-4B	Direct Answer	18.05	12.42	13.40	19.98	16.81	10.58	13.04
InternVL3-8B	Direct Answer	28.25	17.39	20.10	29.99	25.21	19.27	21.74
GLM-4.1V-9B	Direct Answer	25.01	13.66	19.59	26.53	22.69	17.38	13.04
LLaVA-v1.5-7B	Direct Answer	10.93	8.07	7.22	11.35	12.61	7.74	8.70
LLaVA-v1.5-13B	Direct Answer	12.63	4.35	8.25	13.36	13.45	7.27	8.70
LLaVA-Onevision-0.5B	Direct Answer	14.32	12.42	8.25	15.79	14.29	8.37	8.70
LLaVA-Onevision-7B	Direct Answer	21.82	14.91	17.01	23.92	21.01	13.59	13.04
DeepSeek-VL2-3B	Direct Answer	20.37	16.15	15.98	23.13	15.13	11.85	17.39
DeepSeek-VL2-16B	Direct Answer	24.67	21.12	18.04	27.50	19.33	15.96	8.70
DeepSeek-VL2-27B	Direct Answer	26.42	22.36	20.10	28.48	23.53	19.12	13.04
 <i>Proprietary VLMs</i>								
GPT-4o	Direct Answer	43.88	29.19	31.44	44.87	42.86	34.44	30.43
Claude 4 Sonnet	Direct Answer	37.93	29.19	35.57	40.32	37.82	29.07	30.43
Gemini 2.5 Pro	Direct Answer	43.73	32.92	43.81	45.36	44.54	36.18	56.52
 <i>Heuristic RAG</i>								
LLaMA-3.2-11B	Image RAG	18.53	8.07	9.79	19.06	18.49	12.32	4.35
LLaMA-3.2-11B	Text RAG	14.08	8.07	12.37	15.3	14.29	9.16	4.35
LLaMA-3.2-11B	Multimodal RAG	15.19	10.56	13.92	17.06	14.29	9.32	4.35
Qwen2.5-VL-7B	Image RAG	20.66	11.18	13.40	22.71	14.29	12.48	4.35
Qwen2.5-VL-7B	Text RAG	17.56	6.21	12.37	19.25	15.97	9.32	8.70
Qwen2.5-VL-7B	Multimodal RAG	18.53	9.32	9.28	20.64	10.92	9.64	8.70
SUPERLENS[†] (Ours)	Multimodal RAG	38.90 ↑13.69	26.09 ↑12.43	28.87 ↑14.95	40.01 ↑13.66	30.96 ↑7.43	34.45 ↑17.55	26.09 ↑21.74
SUPERLENS[‡] (Ours)	Multimodal RAG	45.23 ↑10.59	37.27 ↑13.67	37.63 ↑16.50	48.76 ↑11.84	33.97 ↑6.24	37.82 ↑14.12	21.74 ↑4.35

Table D. Performance of Different Evaluators on Assessment.

Evaluator	Llama-3.1-8B	Gemma-3-27B	Qwen2.5-32B	Hybrid LLM	Human
Qwen2.5-VL-7B	41.37	39.05	32.82	29.36	32.81
Gemini 2.5 Pro	52.18	52.10	43.02	34.64	43.75
SUPERLENS[‡]	54.75	51.11	44.10	38.03	45.31

Table E. Different Language Performance Comparison.

Language	en (#1369)	zh (#880)	fr (#152)	jp (#63)
Qwen2.5-VL-7B	35.06	32.16	30.26	36.51
Gemini 2.5 Pro	44.63	46.82	37.50	50.79
SUPERLENS[‡]	47.11	44.09	47.37	47.62

and poorly structured query (“American Payroll Association Basic Guide to Payroll author other books”), which conflates the book title with the intent of the question and fails

to isolate the key entity, i.e., the author. As a result, the retriever returns incomplete or misleading evidence. This example indicates that even when the correct tool (hybrid image and text search) is selected, an ill-formed query severely



 <p>Query: What other books has the author of this book written?</p> <p>Reasoning:</p> <ol style="list-style-type: none"> 1. The exact name of the object in the image that the query is about is "American Payroll Association Basic Guide to Payroll". 2. The author of this book is not visible in the image. 3. Without the author's name, I cannot determine what other books the author has written. <p>Therefore, the answer is: {"reasoning": "I cannot determine the author of the book in the image."}</p> <p>Decision: Need RAG Tool Decision: Need Image Search, Need Text Search Search Query: American Payroll Association Basic Guide to Payroll author other books</p>	<p>Object Detection: Book</p>  <p>Search Result:</p> <pre>"html_search": [{"position": 1, "title": "Ok I've more or less been avoiding this app, but this I had ...", "link": "https://www.instagram.com/p/DH-pXZMzQQN/", "source": "Instagram", "source_icon": "https://serpapi.com/searches/690d8516b4ab723cbf81d801", "thumbnail": "https://encrypted-tbn3.gstatic.com/images?q=tbn:~", "thumbnail_width": 251, "thumbnail_height": 201, "image": "https://lookaside.instagram.com/seo/google_widget/crawler/?media_id=", "image_width": 1800, "image_height": 1440}], [{"position": 2, "title": "Had to finally organize half of my ...", ...}]</pre>	<p>RAG Result:</p> <p>Evidence 1 with relevance score 0.86: Show less Read more Like 2 Dislike Reply · 2 replies</p> <p>Evidence 2 with relevance score 0.86: The Bibliophile: For the Love of Books 108 The Bibliophile: For the Love of Books 86 The Bibliophile: For the Love of Books 255 Related Videos</p> <p>Reasoning: The evidence lists multiple titles by the author "The Bibliophile: For the Love of Books," including "The Bibliophile: For the Love of Books 108," "The Bibliophile: For the Love of Books 86," and "The Bibliophile: For the Love of Books 255," indicating that the author has written several books. Answer: The author of "The Bibliophile: For the Love of Books" has written multiple books, including "The Bibliophile: For the Love of Books 108," "The Bibliophile: For the Love of Books 86," and "The Bibliophile: For the Love of Books 255." Ground Truth: Besides this book, Joanne Mitchell-George has also written "The California Payroll Guide".</p>
--	--	---

Figure I. Failure case caused by incorrect search query generation.

degrades retrieval quality.

Together, these two failure cases reveal complementary weaknesses of the system: the first arises from incorrect tool selection, while the second reflects deficient query formulation. These failures emphasize the importance of modality-aware decision-making and precise query construction for building robust multimodal RAG systems.

Case Study in SUPERGLASSES: From Campbell's Soup Can Image to Andy Warhol's American Nationality



Question:	Which country is the renowned artist who painted this item from?
Answer:	<i>Campbell's</i> is painted by American pop-artist Andy Warhol.
Glasses:	Xiao Mi
Image quality:	Normal
Domain:	Food
Location:	Canada
Category:	Multi-hop
Question dynamism:	Static
Difficulty:	Hard
Hops number:	4

Figure J. Case Study in SUPERGLASSES: "Campbell's Soup Can"

Hop 1

Campbell's 22 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The Campbell's Company (doing business as **Campbell's** and formerly known as the **Campbell Soup Company**) is an American company, most closely associated with its flagship **canned soup** products. The classic red-and-white can design used by many Campbell's branded products has become an American icon, and its use in pop art was typified by American artist *Andy Warhol's* series of *Campbell's Soup Cans* prints.

Campbell's has grown to become one of the largest processed food companies in the United States through *mergers and acquisitions*, with a wide variety of products under its flagship Campbell's brand as well as other brands including *Pepperidge Farm*, *Snyder's of Hanover*, *V8*, and *Swanson*. With its namesake brand Campbell's produces soups and other canned foods, baked goods, beverages, and snacks. It is headquartered in Camden, New Jersey.

History



Foundation and early history

The company was started in 1869 by Joseph A. Campbell, a fruit merchant from Bridgeton, New Jersey, and Abraham Anderson, an icebox manufacturer from South Jersey.^[1] They produced canned tomatoes, vegetables, jellies, soups, condiments, and minced meats.

In 1876, Anderson left the partnership and the company became the "Joseph A. Campbell Preserve Company."^[2] Anderson's son, Campbell Spelman, split paths with his father and continued to work at Campbell's as a creative director, originally designing the Campbell's soup cans.^[3]

In 1894, Campbell retired and Arthur Dorrance became the company president.^[4] Campbell reorganized into "Joseph Campbell & Co." in 1896. In 1897, John T. Dorrance, a nephew of company president Dorrance, began working for the company at a wage of \$7.50 a week (\$253 in 2022 dollars).^{[4][5]} Dorrance, a chemist with degrees from MIT and Göttingen University, Germany, developed a commercially viable method for condensing soup by halving the quantity of its heaviest ingredient: water.^[1] He

The Campbell's Company



Trade name: Campbell's
Formerly: Anderson & Campbell (1869–1876), Joseph A. Campbell Preserve Company (1876–1896), Joseph Campbell & Co. (1896–1922), Campbell Soup Company (1922–2024)
Company type: Public
Traded as: Nasdaq: CPB, S&P 500 component, NYSE: CPB (until 2024)
Industry: Food processing
Founded: 1869, 156 years ago
Founder: Joseph A. Campbell
Headquarters: Camden, New Jersey, U.S.
Key people: Mick Beekhuizen (president and CEO),^[1] Keith R. McLoughlin (chairman)
Products: Campbell's, Pepperidge Farm, Pace, V8
Revenue: ▲ US\$10.3 billion (2023)^[1]
Profit: ▲ HK\$ 1.7 billion

Sub-question: What is this product?

Tool used: Image Search

Search URL: <https://en.wikipedia.org/wiki/Campbell%27s>

Search result (snippet): *The Campbell's Company* (doing business as **Campbell's**) . . . its flagship **canned soup** products.

Hop 2

Campbell's Soup Cans 22 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

*This article is about the artwork by Andy Warhol. For other uses, see *Campbell's Soup Cans (disambiguation)*.*


Campbell's Soup Cans is a series of 32 paintings produced between November 1961 and June 1962^{[1][2]} by the American pop art artist Andy Warhol. Each canvas measures 20 inches (51 cm) in height and 16 inches (41 cm) in width and contains a painting of a Campbell's Soup can.^[3] The works were Warhol's hand-painted depictions of printed imagery deriving from commercial products and *popular culture* and belong to the pop art movement.

Warhol began as commercial *illustrator* in 1949. The series was first shown on July 9, 1962,^{[4][5]} at the Ferus Gallery in Los Angeles, California. The exhibition marked the West Coast debut of pop art.^[6] Blum owned the paintings until he loaned it to the National Gallery of Art for several years in 1987 and then sold it to the Museum of Modern Art in 1996. Warhol's motives as an artist were questioned, but the work has become embraced as the most transformative work of art in terms of reconsidering the meaning of art since Marcel Duchamp's 1917 piece *Fountain*. Warhol's association with the subject led to his name becoming synonymous with the *Campbell's Soup Can* paintings.

The Campbell Soup Company was offended at first and considered litigation but soon embraced Warhol's imagery. He eventually produced numerous reproductions of the cans across three distinct phases of his career, while also creating many other works depicting the visual language of commerce and mass media. The soup cans series is generally thought of as referring to the original 32 canvases, but also his many other productions: some 20 similar Campbell's Soup painting variations were also made in the early 1960s; 20 3 feet (91 cm) in height × 2 feet (61 cm) in width, multi-colored canvases from 1965; related Campbell's Soup drawings, sketches, and stencils over the years; two different 250-count 10-element sets of screen prints produced in 1968 and 1969; and other inverted/reversed Campbell's Soup can painting variations in the 1970s. Because of the eventual popularity of the entire series of similarly themed works, Warhol's reputation grew to the point where he was not only the most-renowned American pop-art artist,^[7] but also the highest-priced living American artist.^[8]

The later screen print sets are sometimes confused as part of the original series. In addition, there is ongoing production and sale of unauthorized screen prints, of what is legally Warhol's intellectual property, as a result of a falling out with former employees. The series has a continuing legacy in pop culture, in derivative work by other artists and with multi-million dollar sales in the resale market. The popular explanation of his choice of the soup cans theme is that an acquaintance inspired the original series with a suggestion that brought him closer to his roots.

Campbell's Soup Cans



Artist: Andy Warhol
Year: 1962
Catalogue: 78009P
Medium: Synthetic polymer paint on canvas
Dimensions: 20 by 16 inches (51 cm × 41 cm) each for 32 canvases
Location: Museum of Modern Art, Acquired from Irving Blum in 1996, New York (32 canvas series displayed by year of introduction)
Accession: 476.1996.1-32

Sub-question: What is the famous painting depicting this item?

Tool used: Text Search

Search keywords: *What is the famous painting depicting campbell?*

Search URL: https://en.wikipedia.org/wiki/Campbell%27s_Soup_Cans

Search result (snippet): *Campbell's Soup Cans* is a series of 32 paintings (1961–62) by the American pop-art artist Andy Warhol, each canvas depicting a Campbell's soup can and now considered an icon of pop art.

Figure K. Hop 1 and Hop 2 of “Campbell's Soup Can”

Hop 3

Andy Warhol 140 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

"Warhol" redirects here. For other uses, see Warhol (disambiguation) and Andy Warhol (disambiguation).


Andy Warhol (/wɔːˈthɔː/ [ⓘ]; born **Andrew Warhola Jr.**; August 6, 1928 – February 22, 1987) was an American visual artist, film director and producer. A leading figure in the pop art movement, Warhol is generally considered among the most important American artists of the second half of the 20th century.^[a] His works explore the relationship between artistic expression, advertising, and celebrity culture that flourished by the 1960s, and span a variety of media, including painting, sculpture, photography, and filmmaking. Some of his best-known works include the silkscreen paintings *Campbell's Soup Cans* (1962) and *Marilyn Diptych* (1962), the experimental film *Chelsea Girls* (1966), the multimedia events known as the *Exploding Plastic Inevitable* (1966–67), and the erotic film *Blue Movie* (1969) that started the "Golden Age of Porn".^[2]

Born and raised in Pittsburgh in a family of Rusyn immigrants, Warhol initially pursued a successful career as a commercial illustrator in the 1950s. After exhibiting his work in art galleries, he began to receive recognition as an influential and controversial artist in the 1960s. His New York studio, The Factory, became a well-known gathering place that brought together distinguished intellectuals, drag queens, playwrights, bohemian street people, Hollywood celebrities and wealthy patrons.^{[3][4][5]} He directed and produced several underground films starring a collection of personalities known as Warhol superstars, and is credited with inspiring the widely used expression "15 minutes of fame." Warhol managed and produced the experimental rock band the Velvet Underground. Warhol expressed his queer identity through many of his works at a time when homosexuality was actively suppressed in the United States.^{[6][7]}

After surviving an assassination attempt by radical feminist Valerie Solanas in June 1968, Warhol focused on transforming The Factory into a business enterprise.^[8] He founded *Interview* magazine and authored numerous books, including *The Philosophy of Andy Warhol* (1975) and *Popism* (1980). He also hosted the television series *Fashion* (1979–80), *Andy Warhol's TV* (1980–83), and *Andy Warhol's Fifteen Minutes* (1985–87). Warhol died of cardiac arrhythmia, aged 58, after gallbladder surgery in February 1987.

Warhol has been described as the "bellwether of the art market", with several of his works ranking among the most expensive paintings ever sold.^{[9][10]} In 2013, *Silver Car Crash (Double Disaster)* (1963) sold for \$105 million, setting a record for the

Andy Warhol



Warhol in 1980

Born	Andrew Warhola Jr. August 6, 1928 Pittsburgh, Pennsylvania, U.S.
Died	February 22, 1987 (aged 58) New York City, U.S.
Resting place	St. John the Baptist Byzantine Catholic Cemetery, Bethel Park, Pennsylvania
Education	Carnegie Institute of Technology
Known for	Printmaking • painting • cinema • photography
Notable work	<i>Chelsea Girls</i> (1966 film) <i>Exploding Plastic Inevitable</i> (1966 event) <i>Campbell's Soup Cans</i> (1962 painting) <i>Marilyn Diptych</i> (1962 painting)
Style	Pop art • contemporary art
Movement	Pop art

Sub-question: Who painted this painting?

Tool used: Text Search

Search keywords: *Who paint Campbell's Soup Cans?*

Search URL: https://en.wikipedia.org/wiki/Andy_Warhol

Search result (snippet): Andy Warhol was an American visual artist, film-director and producer, widely regarded as a leading figure of the pop-art movement.

Hop 4

Andy Warhol 140 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

"Warhol" redirects here. For other uses, see Warhol (disambiguation) and Andy Warhol (disambiguation).


Andy Warhol (/wɔːˈthɔː/ [ⓘ]; born **Andrew Warhola Jr.**; August 6, 1928 – February 22, 1987) was an American visual artist, film director and producer. A leading figure in the pop art movement, Warhol is generally considered among the most important American artists of the second half of the 20th century.^[a] His works explore the relationship between artistic expression, advertising, and celebrity culture that flourished by the 1960s, and span a variety of media, including painting, sculpture, photography, and filmmaking. Some of his best-known works include the silkscreen paintings *Campbell's Soup Cans* (1962) and *Marilyn Diptych* (1962), the experimental film *Chelsea Girls* (1966), the multimedia events known as the *Exploding Plastic Inevitable* (1966–67), and the erotic film *Blue Movie* (1969) that started the "Golden Age of Porn".^[2]

Born and raised in Pittsburgh in a family of Rusyn immigrants, Warhol initially pursued a successful career as a commercial illustrator in the 1950s. After exhibiting his work in art galleries, he began to receive recognition as an influential and controversial artist in the 1960s. His New York studio, The Factory, became a well-known gathering place that brought together distinguished intellectuals, drag queens, playwrights, bohemian street people, Hollywood celebrities and wealthy patrons.^{[3][4][5]} He directed and produced several underground films starring a collection of personalities known as Warhol superstars, and is credited with inspiring the widely used expression "15 minutes of fame." Warhol managed and produced the experimental rock band the Velvet Underground. Warhol expressed his queer identity through many of his works at a time when homosexuality was actively suppressed in the United States.^{[6][7]}

After surviving an assassination attempt by radical feminist Valerie Solanas in June 1968, Warhol focused on transforming The Factory into a business enterprise.^[8] He founded *Interview* magazine and authored numerous books, including *The Philosophy of Andy Warhol* (1975) and *Popism* (1980). He also hosted the television series *Fashion* (1979–80), *Andy Warhol's TV* (1980–83), and *Andy Warhol's Fifteen Minutes* (1985–87). Warhol died of cardiac arrhythmia, aged 58, after gallbladder surgery in February 1987.

Warhol has been described as the "bellwether of the art market", with several of his works ranking among the most expensive paintings ever sold.^{[9][10]} In 2013, *Silver Car Crash (Double Disaster)* (1963) sold for \$105 million, setting a record for the

Andy Warhol



Warhol in 1980

Born	Andrew Warhola Jr. August 6, 1928 Pittsburgh, Pennsylvania, U.S.
Died	February 22, 1987 (aged 58) New York City, U.S.
Resting place	St. John the Baptist Byzantine Catholic Cemetery, Bethel Park, Pennsylvania
Education	Carnegie Institute of Technology
Known for	Printmaking • painting • cinema • photography
Notable work	<i>Chelsea Girls</i> (1966 film) <i>Exploding Plastic Inevitable</i> (1966 event) <i>Campbell's Soup Cans</i> (1962 painting) <i>Marilyn Diptych</i> (1962 painting)
Style	Pop art • contemporary art
Movement	Pop art

Sub-question: What nationality is he?

Tool used: Text Search

Search keywords: *What nationality is Andy Warhol?*

Search URL: https://en.wikipedia.org/wiki/Andy_Warhol

Search result (snippet): Andy Warhol was an American artist; he is generally considered one of the most influential U.S. figures in 20th-century art.

Figure L. Hop 3 and Hop 4 of "Campbell's Soup Can"

References

- [1] Jina AI. jina-reranker-m0: Multilingual multimodal document reranker, 2025. 4
- [2] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022. 1, 3
- [3] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, 2023. 3
- [4] Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. Seeclck: Harnessing gui grounding for advanced visual gui agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332, 2024. 1
- [5] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488, 2023. 1
- [6] Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S. Yu, and Ranjay Krishna. Seeking and updating with live visual knowledge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 3
- [7] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [8] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, jia-ayi lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MM-Search: Unveiling the potential of large models as multimodal search engines. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [9] Zhuohang Jiang, Pangjing Wu, Ziran Liang, Peter Q Chen, Xu Yuan, Ye Jia, Jiancheng Tu, Chen Li, Peter HF Ng, and Qing Li. Hibench: Benchmarking llms capability on hierarchical structure reasoning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5505–5515, 2025. 1
- [10] Zhuohang Jiang, Pangjing Wu, Xu Yuan, Wenqi Fan, and Li Qing. QA-dragon: Query-aware dynamic RAG system for knowledge-intensive visual question answering. In *2025 KDD Cup Workshop for Multimodal Retrieval Augmented Generation*, 2025. 1
- [11] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, et al. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [12] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36: 22820–22840, 2023. 1
- [13] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3
- [14] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1
- [15] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023. 1
- [16] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11888–11898, 2023. 1
- [17] Chenyu Wang, Weixin Luo, Sixun Dong, Xiaohua Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6678–6687. IEEE, 2025. 1
- [18] Feng Wang, Zesheng Shi, Bo Wang, Nan Wang, and Han Xiao. Readerlm-v2: Small language model for html to markdown and json. *arXiv preprint arXiv:2503.01151*, 2025. 4
- [19] Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, et al. Crag-mm: Multimodal multi-turn comprehensive rag benchmark. *arXiv preprint arXiv:2510.26160*, 2025. 1, 3
- [20] Ziyue Wang, Chi Chen, Peng Li, and Yang Liu. Filling the image information gap for vqa: Prompting large language models to proactively ask questions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 1
- [21] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 1
- [22] Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing llms to search. *arXiv preprint arXiv:2506.20670*, 2025. 1
- [23] Chi Zhang, Zhao Yang, Jiaxuan Liu, Yanda Li, Yucheng Han, Xin Chen, Zebiao Huang, Bin Fu, and Gang Yu. Appagent: Multimodal agents as smartphone users. In *Proceed-*

ings of the 2025 CHI Conference on Human Factors in Computing Systems, pages 1–20, 2025. [1](#)

- [24] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. [1](#)