DATABASE LAB @POSTECH

# From LILaC and FiF to Real-World Multimodal Retrieval

2026.2.12

Presenter: **Joohyung Yun**

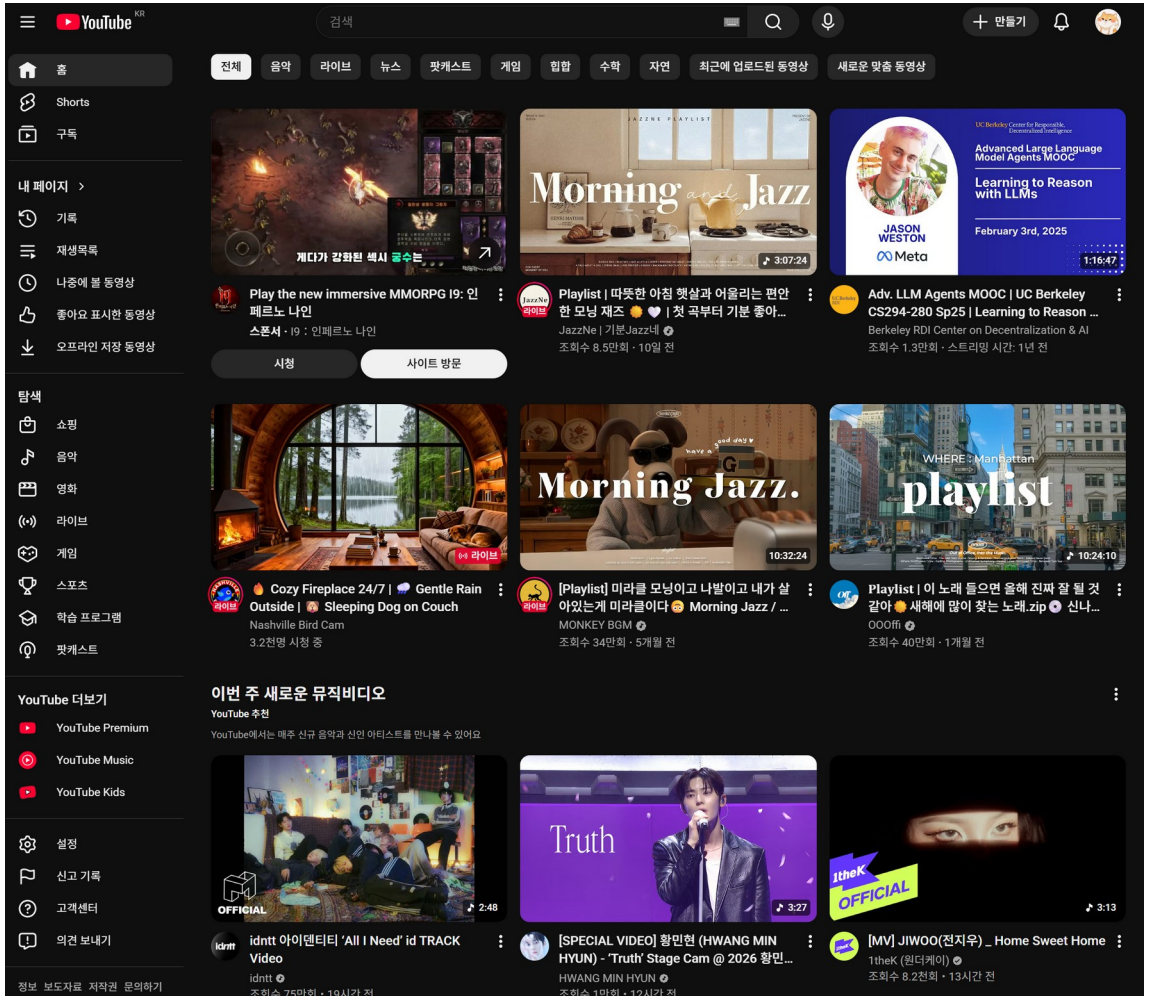
# Overview of This Seminar

- Top Goal of My Research
- Brief Introduction of My Previous Work
- How Does Real-World Multimodal Retrieval Differ
- Future Challenges to Solve
- Viable Research Topic Options

# Retrieving Info From Webpages Matter

- The web is the **largest and most up-to-date knowledge source** — but also the noisiest and most dynamic.
- Next-gen AI products (search, assistants, copilots) require retrieval that works on **real webpages, not curated benchmarks**.
- **Multimodal evidence** is everywhere on the web: text + tables + figures + UI elements + (increasingly) video/audio.

**My long-term goal**: make retrieval robust enough for real-world, open-web data.



# AI Entrepreneurs Want “Web-Native” Retrieval

Goal **Solve information retrieval in the open web at production-grade constraints:**

- Input is raw HTML + rendered pages
- Pages are dynamic/reactive (scroll, expand, login, infinite feed)
- Multi-modality expands (audio/video/UI interactions)
- Tight latency/cost budgets + reliability (provenance, robustness)

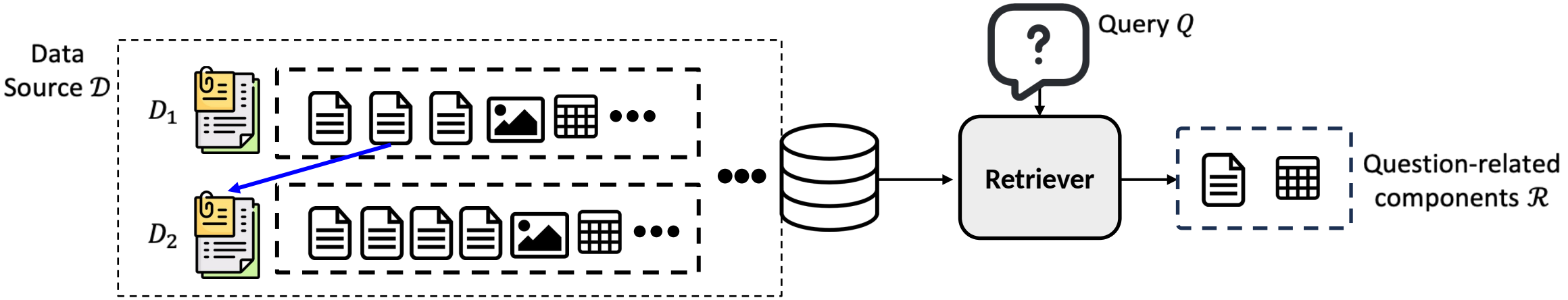
# Brief Introduction of My Previous Work

A Relaxed Problem Setting

# Open-Domain Multimodal Document Retrieval

**Given a data source $\mathcal{D}$, a link mapping $\mathcal{L}$ and a query $Q$, correctly retrieve a ranked list of components $\mathcal{R} = [C_1, \ldots, C_{n_{ret}}]$. $\mathcal{R}$ should contain the ground truth set of components $\{C_{gt_1}, \ldots, C_{gt_r}\}$.**

- Data source $\mathcal{D} = \{D_1, \ldots, D_{k_{doc}}\}$
- Document $D = \{C_1, \ldots, C_{k_{comp}}\}$
- Component $C \in \{P, T, I\}$
  - $P$: paragraph, $T$: table, $I$: image
- Link mapping $\mathcal{L} = \mathcal{C} \rightarrow \mathcal{D}$
  - $\mathcal{C}$: A set of all components within $\mathcal{D}$



# LILaC (2025): Embedding-based Traversal on Graph Structure

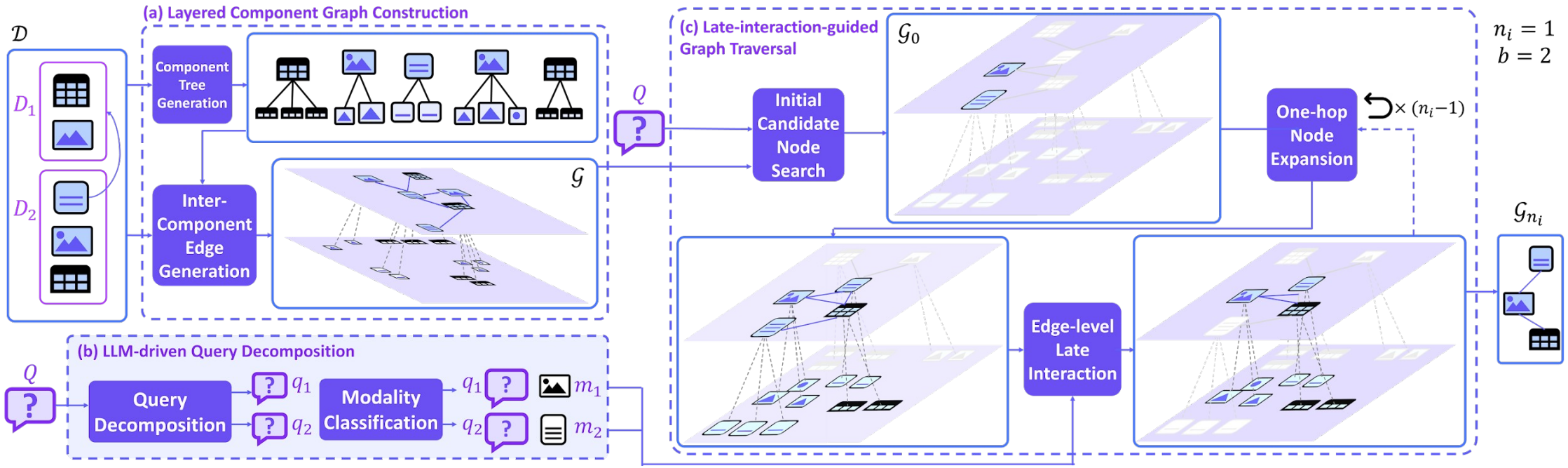
**Core idea** **Organize documents into a 2-layer graph and traverse edges with late interaction scores.**

**Enables**

- Coarse-to-fine evidence
- Multihop via navigational links
- Edge scoring with late interaction

**Key assumptions**

- Offline graph + embeddings
- Pre-extracted components/subcomponents
- Reliable hyperlink mapping



# FiF (2026): Failure-aware, Cost-aware Agentic Traversal

**Core idea**

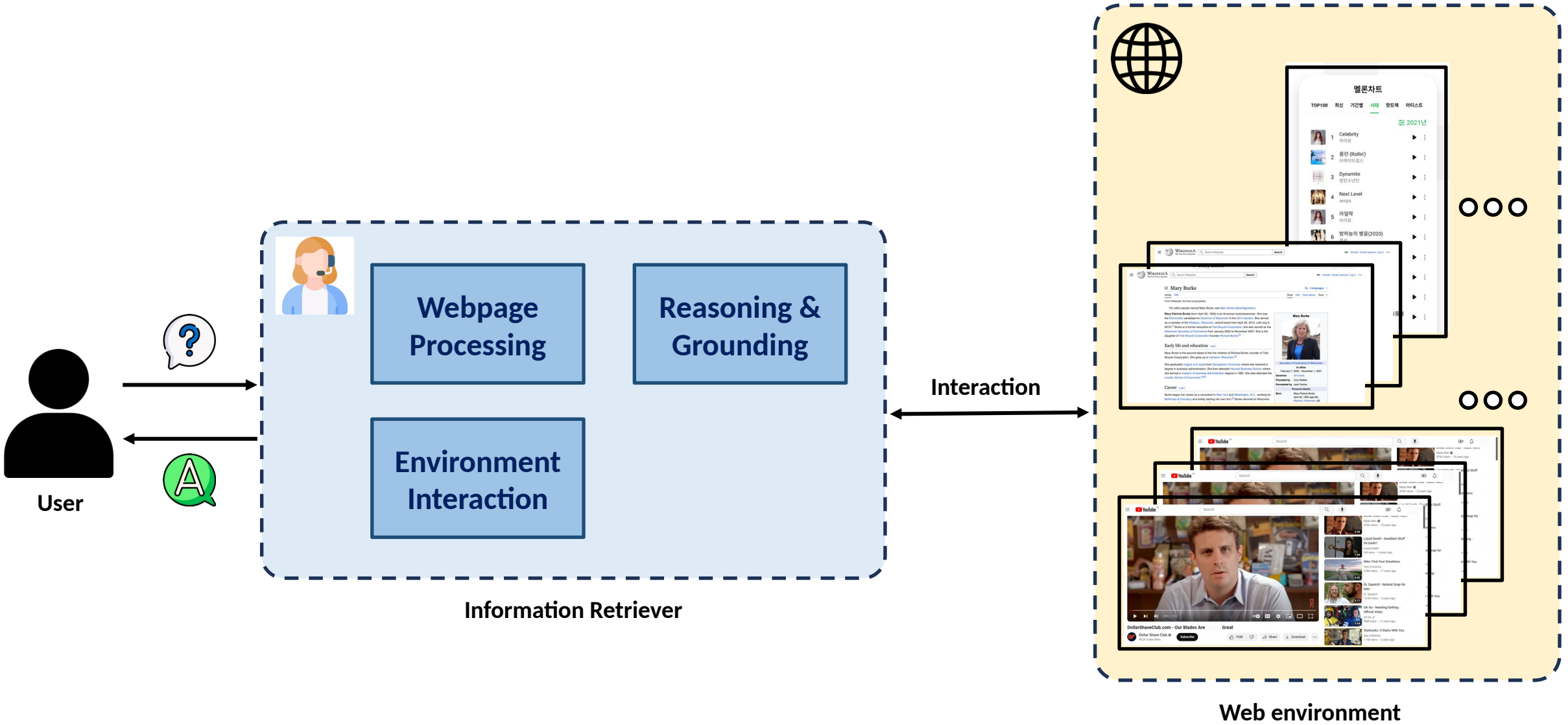
**Treat multihop retrieval as an information-state MDP: revise subqueries, switch strategies, and backtrack using failure traces.**

**Still assumes**

- Clean component graphs + reliable navigational signals
- Static pages (no reactive UI / hidden content)
- Offline corpora with URL annotations

# What Should be Considered Additionally for a Practical Real-World Multimodal Retrieval?

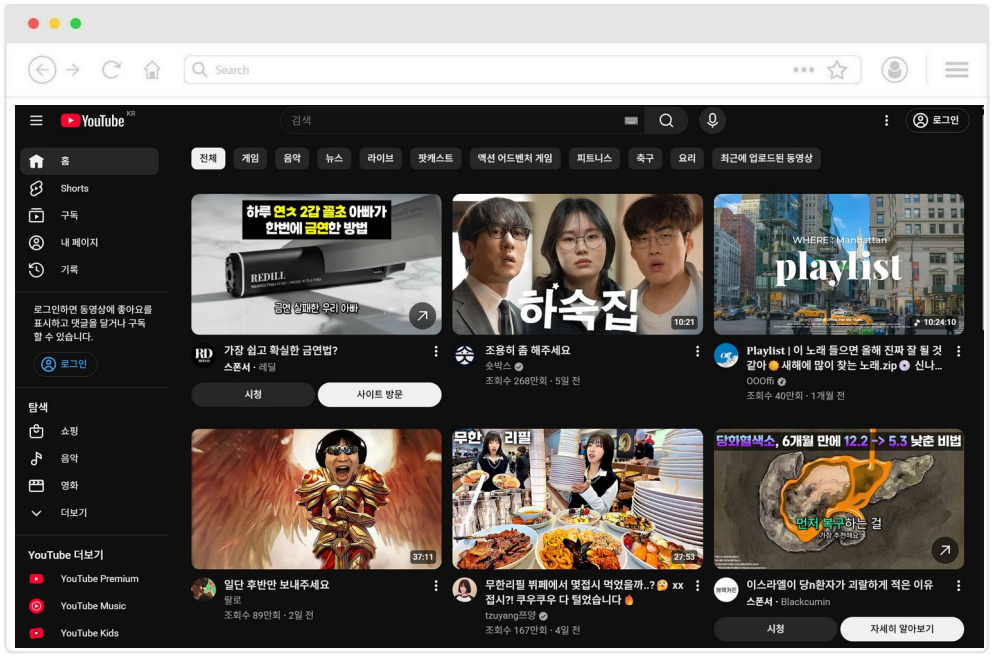
# Diagram of Practical Real-World Information Retriever



# Real Web Input ≠ Preprocessed Components

## The retriever should start from raw web

- Input is raw HTML/DOM + CSS/JS + rendered pixels (not clean paragraphs/tables/images).
- Need end-to-end extraction: boilerplate removal, component segmentation, table parsing, figure/plot understanding.




Raw web

Immediate outputs

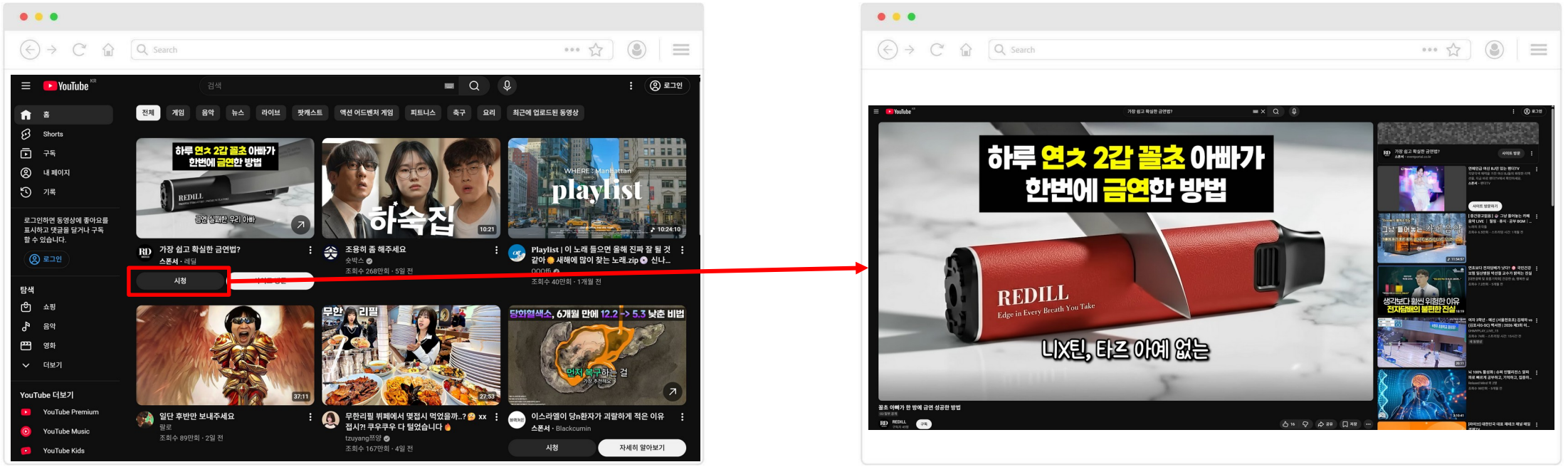
Processed components

# More Modalities: Beyond Text, Table and Image

**Multi-modality**

- Video: retrieve relevant time spans and frames
- Audio: retrieve segments and transcripts.

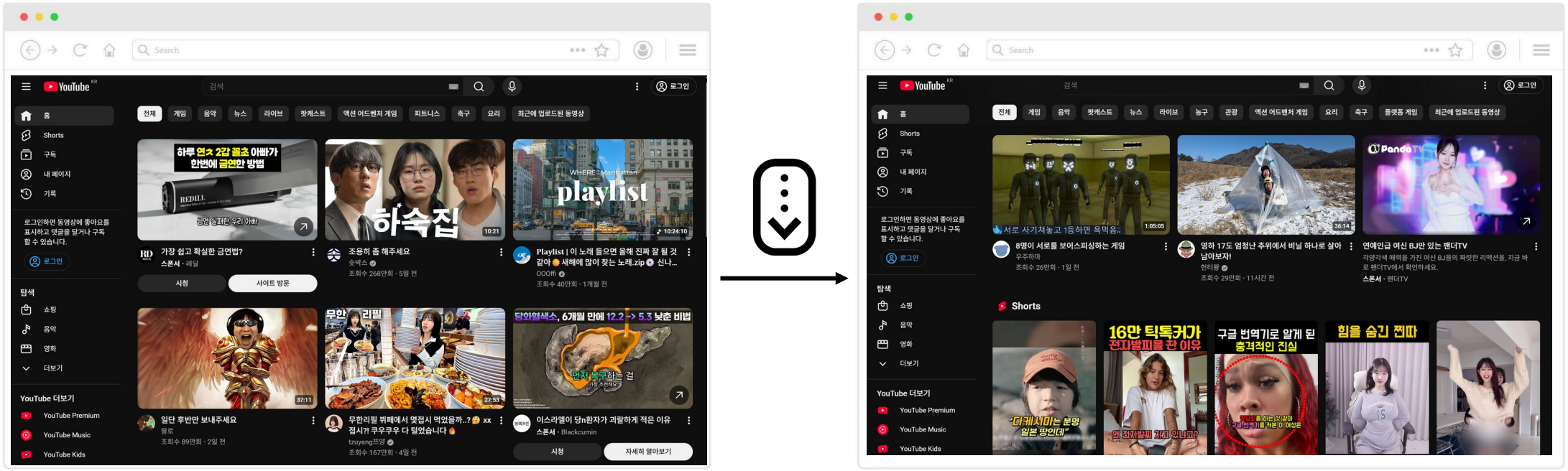
**UI signals**: buttons, menus, tooltips, and layout semantics can carry meaning.



# Dynamic and Reactive Webpages

## Evidence may be hidden behind actions

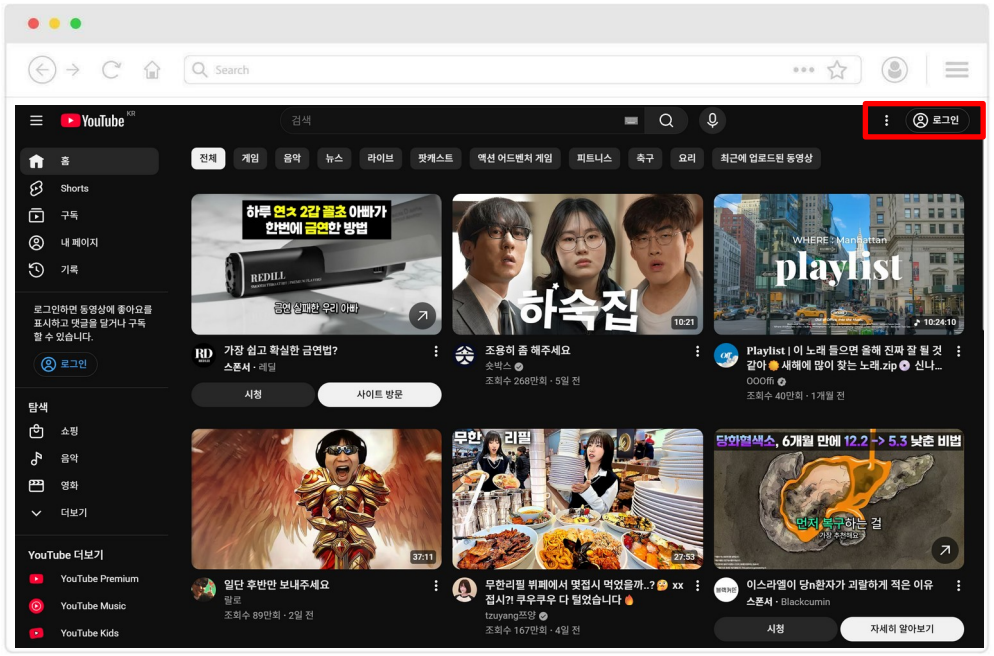
- Content appears after scroll, expand, tab switch, pagination, login, or async requests.
- Retrieval becomes stateful: observe → act → re-observe → retrieve



# Challenges in Offline Indexing

## The corpus changes continuously and is just too big

- Indexing should be done incremental/streaming, for the entire web environment.
- Pages update, disappear, and personalize.



Web environment

# Grounded Component Retrieval

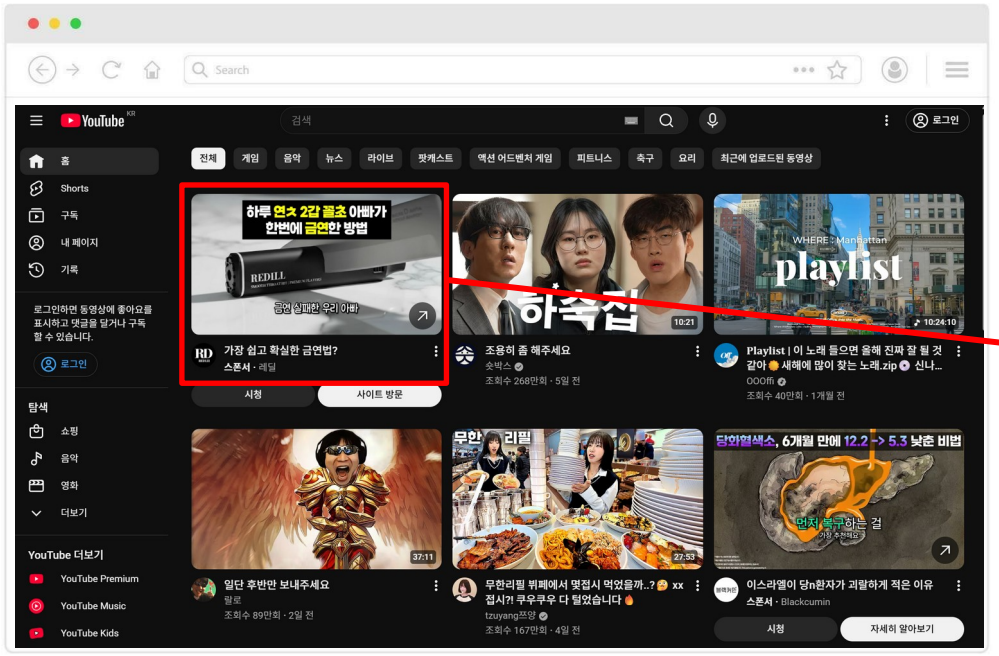
## Retrieve evidence with grounding

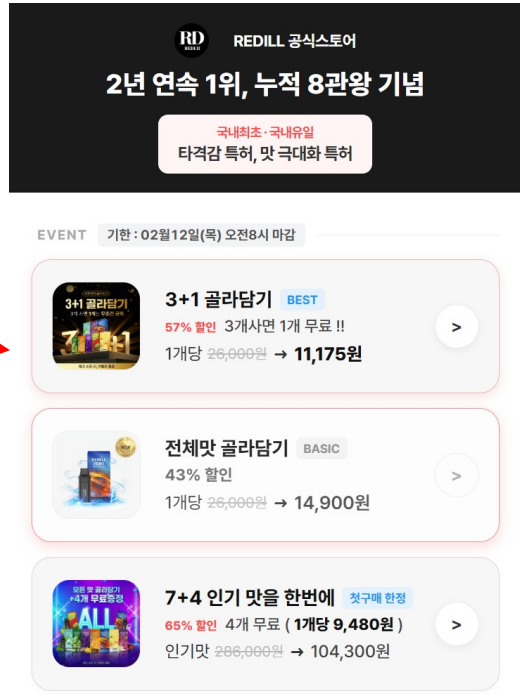
- Handle multimodal grounding: cite the exact table rows / chart region / image crop used.

# Production Constraints

## Enterprise-grade budgets and reliability

- Latency/cost budgets: avoid full-page LLM/OCR per request; use adaptive pipelines and caching.
- Robustness: withstand spam/SEO, templates, duplicates, and noisy link signals.
- Compliance: robots.txt, access control, PII, copyright; must support safe failure modes.



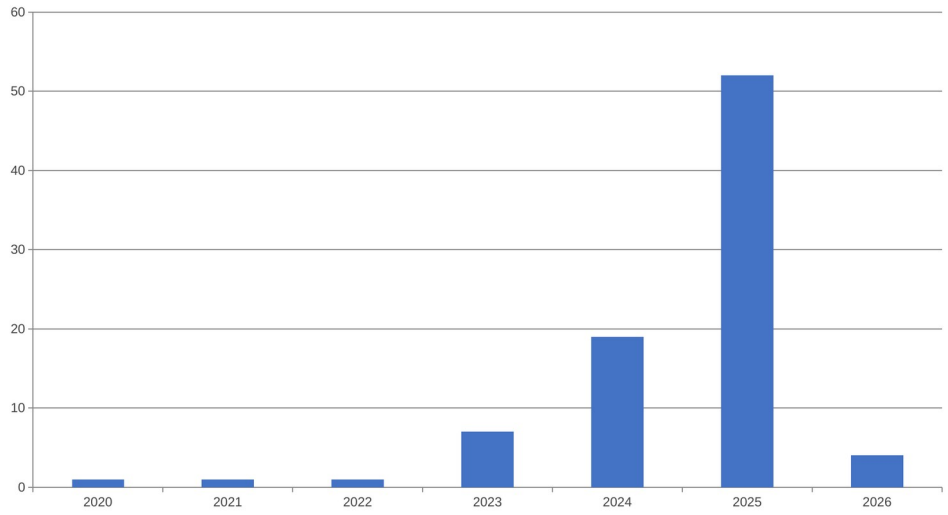



# Current Work & Their Challenges in Solving Real-World Multimodal Retrieval

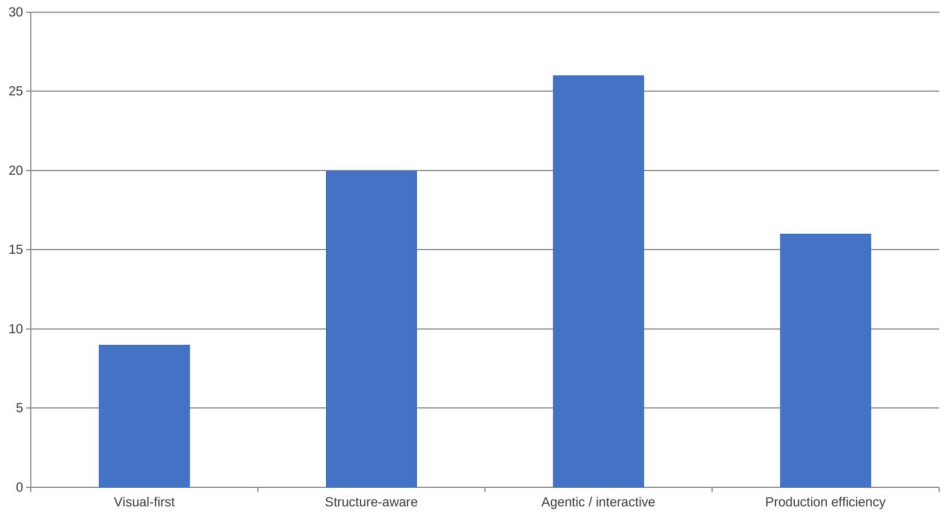
# How the Research Community Tackles “Web Realism”

*Survey: 85 papers (2020–2026). Tags are non-exclusive (a paper can belong to multiple directions).*

**Papers Published per Year**



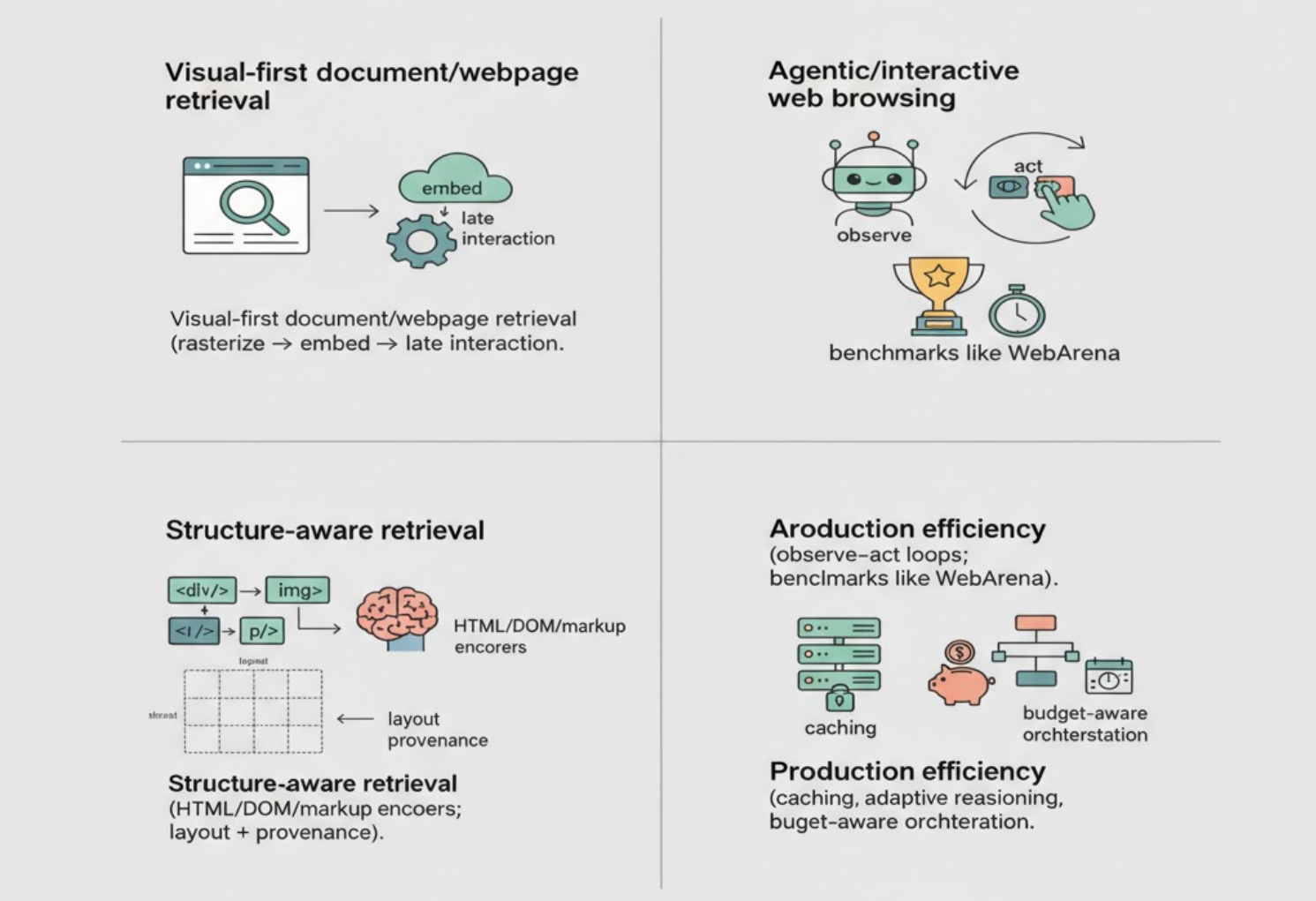

**Papers Published for Each Direction**




| Venue | # papers | Share |
|---|---|---|
| arXiv | 43 | 50.6% |
| NeurIPS / ICLR / ICML | 18 | 21.2% |
| ACL / EMNLP / NAACL | 16 | 18.8% |
| ACM MM | 5 | 5.9% |
| SIGIR / TPAMI | 2 | 2.4% |
| Other | 1 | 1.2% |

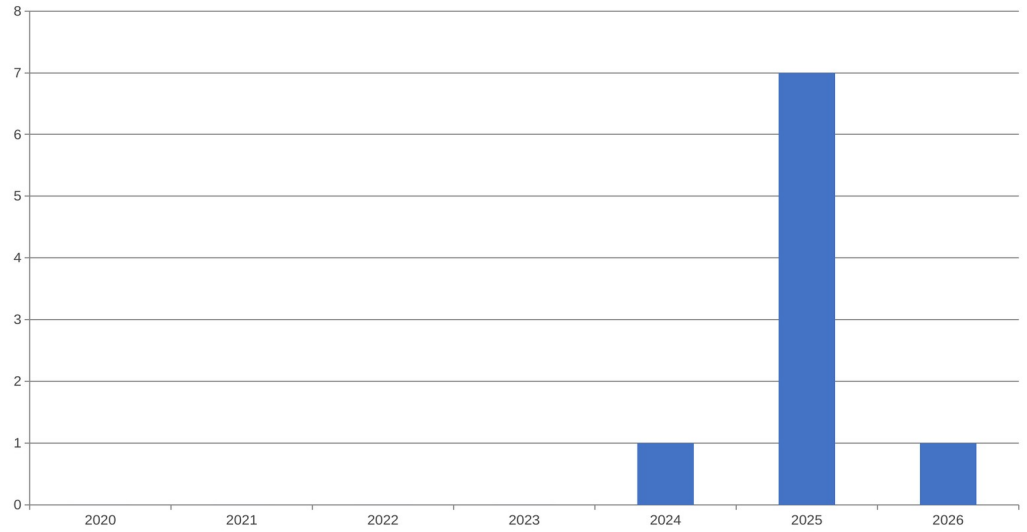
# How the Research Community Tackles “Web Realism”

## Four directions emerging across ML/NLP + industry



# Visual-Embedding Retrieval

Papers tagged
**9**




| Venue | # | Share |
|---|---|---|
| arXiv | 6 | 67% |
| NeurIPS / ICLR / ICML | 1 | 11% |
| ACL / EMNLP / NAACL | 1 | 11% |
| ACM MM | 1 | 11% |

## Typical approach

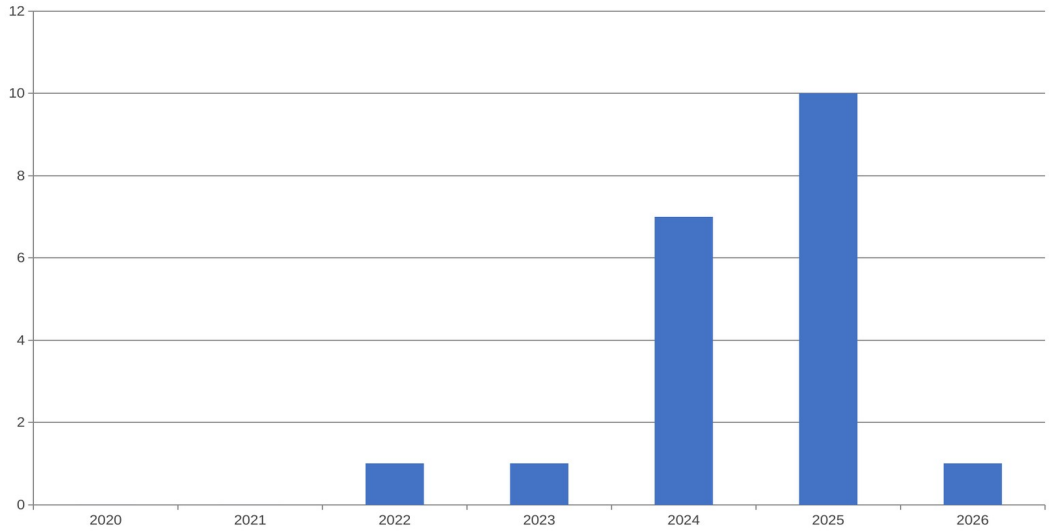
- Rasterize webpages (full page / regions) into VLM embeddings.
- Exploit layout as a signal, often reducing dependence on brittle HTML parsing.
- Use OCR/region proposals when grounding is required.

## Open gaps

- Region alignment: align pixel regions back to DOM nodes and page states.
- Cost: rendering/OCR is expensive; need caching + on-demand execution.
- Robustness to templates, ads, and responsive layouts.

# Structure-Aware Retrieval

Papers tagged
**20**




| Venue | # | Share |
|---|---|---|
| arXiv | 10 | 50% |
| NeurIPS / ICLR / ICML | 6 | 30% |
| ACL / EMNLP / NAACL | 2 | 10% |
| ACM MM | 1 | 5% |
| SIGIR / TPAMI | 1 | 5% |

## Typical approach

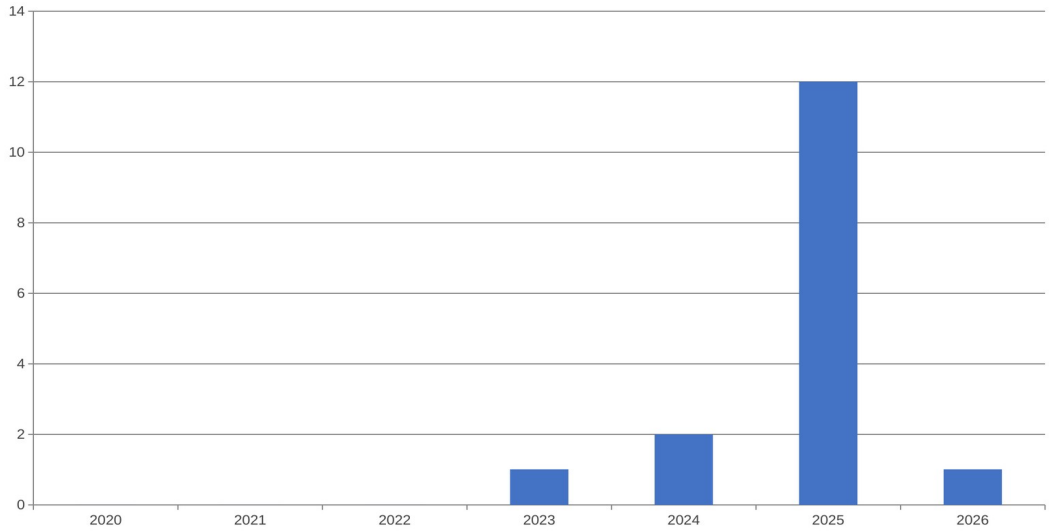
- Encode HTML/DOM/markup with structure-aware Transformers; index blocks/nodes as retrieval units.
- Use DOM hierarchy, attributes, and link structure to complement text matching.
- Fuse structure with visual/layout cues for better grounding.

## Open gaps

- End-to-end segmentation under noisy templates & JS-generated DOM.
- Joint modeling of DOM + pixels + cross-page navigation signals.
- Standardized evaluation for grounded evidence and dynamic behavior.

# Production Efficiency

Papers tagged
**16**




| Venue bucket | # | Share |
|---|---|---|
| arXiv | 9 | 56% |
| NeurIPS / ICLR / ICML | 3 | 19% |
| ACL / EMNLP / NAACL | 2 | 12% |
| SIGIR / TPAMI | 1 | 6% |
| Others | 1 | 6% |

## Typical approach

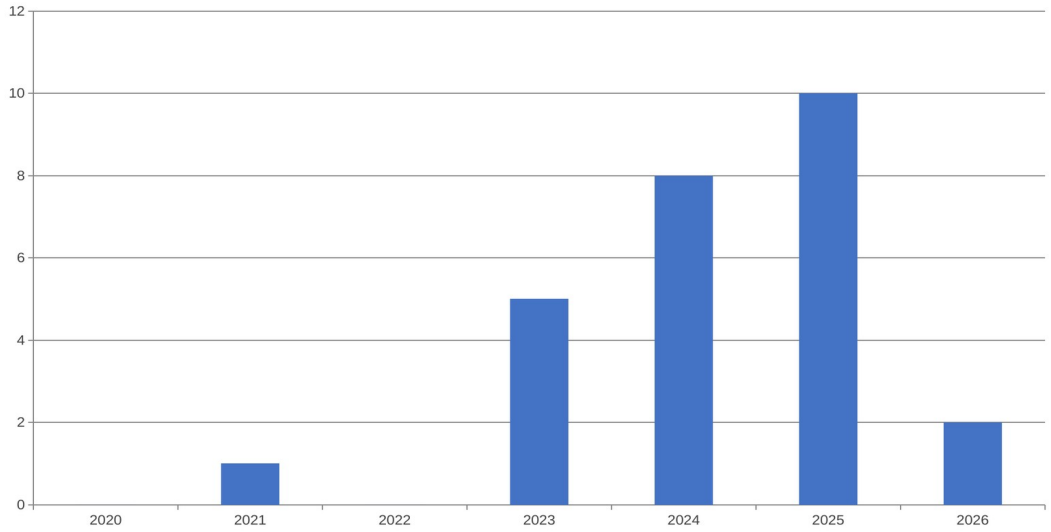
- Caching (embeddings, KV-cache), adaptive compute, and selective recomputation.
- Streaming updates + freshness-aware ranking for changing corpora.
- System-level evaluation that includes latency and cost.

## Open gaps

- Unified optimization across retrieval + generation + browsing actions under SLOs.
- Benchmarks for realistic workloads (long tail, partial failures, distribution shifts).
- Compliance/safety constraints integrated into end-to-end pipelines.

# Agentic Web Browsing

Papers tagged
**26**




| Venue | # | Share |
|---|---|---|
| arXiv | 11 | 42% |
| NeurIPS / ICLR / ICML | 7 | 27% |
| ACL / EMNLP / NAACL | 6 | 23% |
| ACM MM | 2 | 8% |

## Typical approach

- Formulate retrieval as sequential decision-making in browser environments (observe–act–retrieve).
- Observation pruning / line retrieval to control context and costs.
- Backtracking and tool use to recover from dead-ends.

## Open gaps

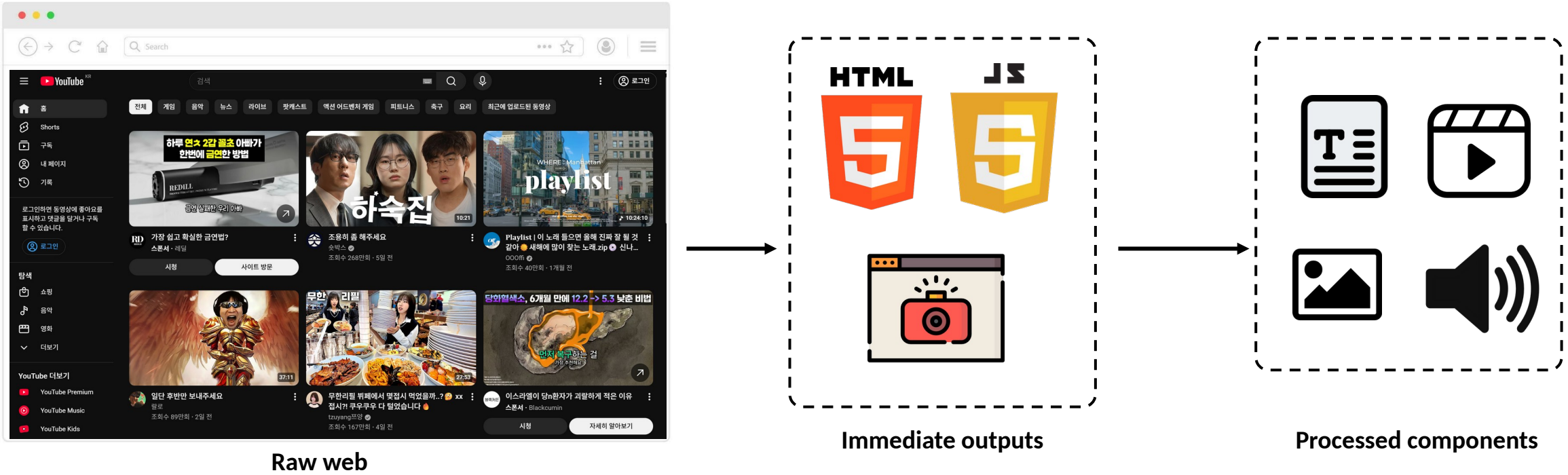
- Multimodal evidence retrieval in the loop (tables/figures/video) with grounding.
- Budgeted control (latency/cost) with reliability guarantees.
- Generalization across sites + adversarial/SEO noise.

# What Should We Research Next?

# Topic 1: Web-Native Processing Within Retrieval

## End-to-end: HTML/DOM + render → components → grounded retrieval

- Goal: remove the “pre-parsed components” assumption in LILaC/FiF.
- Key research: robust componentization under noise + provenance-aware ranking.



# Topic 2: Reactive Web Retrieval

**Retrieval + actions on dynamic pages (observe–act–retrieve)**

- Extend FiF's tool list to atomic real web actions (scroll/click/login).
- Learn when to explore vs. exploit, and when to backtrack or re-anchor.

# Topic 3: Modality Extension

## Extend components to video/audio/temporal spans

- Retrieve: video time ranges, keyframes, audio segments, UI-state evidence.
- Need temporal grounding + scalable indexing for long media.

# Topic 4: Freshness-Aware Web IR under Budgets

**Streaming indexing + caching + adaptive computation**

- Move from offline embedding to hybrid (precompute + on-demand) pipelines.
- Optimize accuracy–freshness–cost jointly (enterprise SLOs).
- Deliverables: retriever with incremental updates and cache-aware routing.