GRAIL-V: Grounded Retrieval & Agentic Intelligence for Vision-Language

Amit Agarwal[1] Vivek Gupta[2] Vivek Srikumar[3] Tao Sheng[1]
Alice Oh[4] Sara Hooker[5] Jyotika Singh[1] Hitesh Laxmichand Patel[6]

[1]Oracle AI [2]Arizona State University [3]University of Utah [4]KAIST [5]Adaption Labs

Abstract

GRAIL-V is a CVPR workshop dedicated to grounded multimodal retrieval and reranking for agentic-vision systems. As agents plan, retrieve, and verify before acting across diverse structured and unstructured data sources, robust evidence grounding and efficiency become bottlenecks for reliable deployment. GRAIL-V brings together CV, IR, NLP, and HCI communities to advance unified methods and evaluation practices for visual–text search, multimodal tool use, and calibrated decision-making. GRAIL-V features keynotes, presentations, and a panel, emphasizing real-world constraints. By fostering shared evaluation protocols and cross-domain collaboration, GRAIL-V aims to catalyze progress toward robust, accountable, and deployable agentic-vision systems, engaging both researchers and practitioners.

GRAIL-V: Grounded Retrieval & Agentic Intelligence for Vision-Language

1. Tracks & Fit to CVPR

Primary track: Vision, language, and reasoning. **Secondary tracks:** Multimodal learning; Recognition: categorization, detection, retrieval; Vision applications and systems; Document analysis and understanding; Foundation Models.

Recent CVPR/ICCV/ECCV papers show rapid, but fragmented, progress at each stage of this loop: *retrieval* frameworks that unify tasks and modalities (LamRA for universal retrieval/reranking; GENIUS for generative multimodal search) [9, 13]; list-wise, long-context *reranking* that reasons over entire shortlists (LOCORE) [20]; and stronger *verification/grounding* at object-, pixel-, and video-tube levels (GLIP; X-Decoder; VideoGrounding-DINO) [10, 19, 21]. For text-rich media, document- and layout-aware LVLM methods (LayoutLLM; DoCo) and robustness studies (RoDLA) push toward faithful answers on pages, tables, and charts [5, 11, 14], while new resources such as MMDocIR, ViDoRe v2, and ChartQAPro expand evaluation difficulty and breadth [7, 16, 18]. On the agent side, visual-planning and API tool use for spatial/UI reasoning (e.g., dynamic visual APIs) is entering CVPR itself [17]. Collectively, these advances highlight the need for *shared, modality-spanning practices* to measure retrieval quality, re-ranker fidelity, groundedness, and end-to-end efficiency under real constraints (latency, memory, cost).

CVPR uniquely concentrates the communities that build vision-grounded retrieval, train re-rankers, and design grounded verification and system benchmarks. A workshop dedicated to the agentic vision systems that *P(lan)→R(etrieve)→R(erank)→V(erify)* is squarely in scope for CVPR's evaluation-focused culture and its call for high-impact systems and benchmarks [8].

2. Motivation and Significance

Modern vision systems are evolving into *agentic* entities that do more than simply return answers: they *plan*, *retrieve*, *rerank*, and *verify* evidence before taking action. This paradigm shift is driven by real-world demands—ranging from medical document triage and legal discovery, to dashboard analysis, chart extraction, and accessible technology. Concretely, such agents (1) plan to break the tasks across steps and tools (2) recall candidate pages, frames, or regions across a spectrum of unstructured media (including images, video, documents, charts, and UI screenshots) and structured sources (such as tables, SQL databases, code, and logs); (3) rerank by jointly reasoning over (query, context, candidate) at region, page, or moment granular-

ity; and (4) verify consistency before returning outputs or invoking the next appropriate tool. Centering this plan-retrieval–reranking–verification loop is essential to maximize reliability—before any action is taken.

In practical deployments, these systems often interleave long PDFs with embedded figures and tables, dashboards and app/web screenshots, time-aligned video and ASR output, as well as schemas, queries, and code—all handled by specialized tools invoked inside the decision loop (e.g., layout/figure detectors, chart/table readers, OCR-free/UI parsers, SQL/DB/code tools). Unifying retrieval and reranking over such mixed-modality, mixed-structure corpora necessitates spatial and temporal grounding—over precise regions, pages, and moments, rather than simple whole-document matches. Achieving effective grounding, calibration, and efficiency under real-world constraints is thus a uniquely vision challenge—one that sits at the intersection of computer vision (CV), information retrieval (IR), natural language processing (NLP), and human computer interaction (HCI). Despite progress, critical research and deployment gaps remain:

- **Hybrid retrieval across structured & unstructured sources.** Unified search over pages, charts, and screenshots *and* SQL/code/metadata with tool-assisted parsing and reasoning.
- **Grounded evaluation beyond Recall@K.** Region/page/moment provenance, *citation fidelity*, and *calibration/abstention* at top-k.
- **Efficiency under deployment constraints.** Throughput, latency, memory, and resource costs as primary metrics for retrievers and rerankers, alongside traditional accuracy.
- **Robustness & Safety.** Defenses against spurious-cue sensitivity, UI/theme drift, prompt/format variation, screenshot/DOM leakage, and tool misuse.
- **Reproducibility & Artifacts.** Standardized overlays, spans, moments, and executable traces to enable apples-to-apples diagnosis and benchmarking.
- **Fragmentation.** Lightweight, *shared* harnesses and community-aligned reporting *practices* to connect currently siloed efforts—without enforcing a prescriptive checklist.

Our long-term vision is that GRAIL-V acts as a launchpad for persistent cross-community infrastructures: collaborative leaderboards, reproducible artifacts, and new interdisciplinary partnerships that extend beyond the workshop itself. Success will be measured by the adoption of shared resources, the establishment of open challenges, and the creation of bridges between fragmented subfields—accelerating the development and deployment of robust, accountable, and impactful agentic vision systems.

3. Topics of Interest

This workshop seeks *vision-first* advances that make agentic planning → retrieval → reranking → verification practical across unstructured imagery/video, and structured sources, with grounded, efficient, and safe evaluation. Topics include (but are not limited to):

Heterogeneous & multimodal retrieval
- Unified visual+text retrieval over images, pages/tables/charts/diagrams, video (incl. temporal/multi-hop), and UI/screenshots.
- Late/early interaction retrievers; hybrid dense+lexical; multilingual and cross-domain retrieval; long-context for visually rich documents.

Multimodal Reranking
- Cross-encoders and efficient late-interaction variants for page/frame/region selection; reasoning-augmented rescoring.
- Throughput/latency trade-offs, batching/caching/distillation; on-device/streaming; stress-testing spurious cues and UI/theme changes.
- Calibration and abstention at top-k; selective prediction for agent hand-offs.

Agentic Planning & Tool Use
- Query reformulation, routing, and composition of tools (OCR-free parsers, table/chart readers, layout/figure detectors, math/code/DB tools).
- *Collaborative editors and IDEs*: retrieval/reranking for co-authoring (docs, wikis, dashboards) and in-IDE assistants grounded error/screenshot/UI evidence.
- Safety-conscious tool use: guardrails, constraints, telemetry, and error recovery.

Grounding, Provenance & Reliability
- Region/span/moment overlays; *citation fidelity* scoring; structured evidence alignment (tables/charts/UI/DOM).
- Consistency checks and verification compatible with agentic pipelines; leakage-resistant protocols for screenshots/forms/interleaved media.

Data enrichment
- Layout/structure extraction, chart metadata, temporal alignment, multimodal descriptors;

- Reproducible *evaluation harnesses*, prompts/traces, and lightweight leaderboards; toolkit integrations (e.g., reranking suites).

4. Relation to Prior Workshops

Recent workshops at CVPR/ICCV/ECCV have emphasized foundation models or broad multimodal applications rather than the mechanics and *evaluation* of evidence acquisition for action: *DriveX* focuses on foundation models for V2X cooperative driving [6]; *Wild3D* on 3D/4D modeling and generation [4, 15]; *GeoLME* on geometric computing with large models [12]; *TAVI* highlighted tool-augmented vision and retrieval-augmented models but did not address grounded re-ranking/verification or run a CFP [3]; *MULA* surveys general multimodal learning and data fusion [2]; and *O-DRUM* targeted open-domain multimodal retrieval without agentic/re-ranking/verification pipelines [1].

GRAIL-V is complementary and non-overlapping: it centers the agentic *plan* → *retrieve* → *re-rank* → *verify* loop and its cross-modality evaluation, unifying stuctured and unstructured data soruces (e.g. region/page/moment across images, video, documents), and interiactive GUI, to promote calibrated-decision making with accepted cost/latency trade-offs, aiming to take research to industry deployments.

Willingness to co-track. If closely related proposals are submitted, we will coordinate to (i) merge or co-track complementary programs and (ii) harmonize CFPs, artifact expectations, and schedules to minimize overlap and maximize community value, consistent with CVPR-26's guidelines [8].

5. Broader Impact

Cross-community Bridge. This workshop connects CVPR's vision community with adjacent areas in information retrieval (embedding+ranking), data management (hybrid structured+unstructured data source), human–computer interaction (UI/agent co-pilots), natural language processing (LLMs/VLMs/agents), and systems (latency/cost/energy). Standardizing *region/page/moment-grounded* evaluation and artifact practices enables reproducible comparisons across labs, toolchains, and deployment settings, and streamlines technology transfer from research to production.

Public value and Safety. Agentic vision systems already support accessibility (region-linked evidence for screen readers), healthcare and finance (licensing-aware document/diagram retrieval with provenance), and education (evidence-backed tutoring over figures/charts). Emphasizing *calibration/abstention*, *provenance*, and *efficiency* improves trust, reduces operational cost/energy, and mitigates harms from spurious cues, leakage, or tool misuse.

6. Speakers

We are excited to invite a diverse range of confirmed speakers and panelists, each bringing significant expertise in multimodal retrieval, vision-language models, and agentic systems.

Kristen Grauman | | University of Texas, Austin

Professor of Computer Science at UT Austin and leader of the UT-Austin Computer Vision Group, Kristen Grauman's research advances video understanding, egocentric perception, and embodied AI—spanning long-form video, audio–visual learning, and first-person interaction. She initiated and technically led the large-scale *Ego4D* and *Ego-Exo4D* international efforts that set new benchmarks and datasets for first-person perception. Her recognitions include IEEE Fellow, AAAI Fellow, the IJCAI Computers & Thought Award, the PECASE, and the CVPR Helmholtz (test-of-time) Prize. At GRAIL-V, her talk will connect video to the agentic *plan* → *retrieve* → *re-rank* → *verify* loop, focusing on long-horizon retrieval, moment-level grounding, and evaluation protocols for calibrated, efficient assistants.

Mohit Bansal | | University of North Carolina, Chapel Hill | *Confirmed*

Mohit Bansal is a leader in multimodal generative models and vision-language reasoning, specializing in efficient and generalizable deep learning methods. As an AAAI Fellow, his work on agentic systems and faithful, controllable generation for vision-language tasks aligns perfectly with the workshop's goal to build robust agentic systems capable of multimodal reasoning.

Yunyao Li | | Adobe | *Confirmed*

Yunyao Li is an ACM Distinguished Member with significant contributions to knowledge graph systems in enterprise settings, including retrieval and QA systems. Her expertise in graph-based retrieval and content aggregation from multimodal sources will add depth to discussions on scalable, efficient multimodal retrieval and tool-assisted reranking.

Dan Roth | | University of Pennsylvania; Oracle AI | *Confirmed*

Dan Roth is a pioneer in knowledge-intensive NLP and learning/inference for language. A Fellow of AAAS, ACM, AAAI, and ACL, he has contributed decades of work on *grounding, reasoning, and robust decision-making*. His leadership in large-scale multimodal re-

trieval and agentic systems for robust inference aligns closely with the workshop's focus on vision-centric retrieval and agentic pipelines, especially for decision-making tasks across multimodal sources.

Nil Reimers | | Cohere | *Confirmed*
Nil Reimers is a leader in dense vector spaces, multimodal embeddings, and retrieval-based content aggregation. His work on bridging vision and language through deep embeddings for search and content retrieval makes him a perfect fit for this workshop, where multimodal retrievers and rerankers will be a key focus.
Iryna Gurevych | | Technical University of Darmstadt | *Confirmed*
Iryna Gurevych is a professor and the Director of the UKP Lab, with expertise in large-scale NLP resources, evaluation methodology, and knowledge-centric NLP. As an ACL Fellow and 2023 ACL President, her work on multilingual embeddings and large-scale multimodal resources will inform the workshop's discussions on evaluation methodologies for vision-language systems, as well as retrieval techniques across diverse languages and domains.
Scott Wen-Tau Yih | | Meta | *Confirmed*
Scott Yih is a Senior Researcher at Meta focusing on semantic parsing, multi-hop question answering, and retrieval-centric evaluation at scale. His work on efficient retrieval models and reasoning over multimodal data is critical to the workshop's focus on agentic systems for vision-language tasks, including tools for multimodal parsing and reranking.

Tentative Panel

Panel on: *From Retrieval to Action: What Should Agentic Vision Systems Verify?*
The panel will feature experts from industry and academia discussing key open questions in multimodal retrieval, the challenges of grounding and verification, and the societal and safety considerations of deploying agentic systems. Panelists will be selected to provide diverse perspectives across *deployment-first vs benchmark-first* approaches, *OCR-free vs OCR-assisted parsing*, and *monolithic vs tool-augmented architectures*.

Notes: All invited speakers are independent, and conflicts of interest (COIs) have been avoided. The funding for this workshop is self-supported, with sponsorship from Oracle currently under discussion.

7. Format & Logistics

Full-Day Request We request a **full-day** workshop (09:00–17:00). The program includes four keynotes, two oral sessions, two posters/demos blocks, a contrarian panel, and a short closing with awards. A half-day format would force either (i) removing the posters/demos engagement or (ii) compressing keynotes and panel, both of which materially reduce community value. The full-day schedule enables balanced depth (methods, systems, evaluation) and breadth (vision+text, structured+unstructured, agentic tool use).

Expected Audience Size: Medium (100–300) in-person (plus hybrid attendance). **Profiles:** vision-language, document/diagram/charts understanding, retrieval/ranking, datasets/evaluation, systems and efficiency, HCI/agent tooling, and applied teams deploying production pipelines. **Submissions:** We expect approx 100 submissions to the workshop.

Hybrid Plan CVPR-provided Zoom room for remote keynotes/panelists as needed; live Q&A bridged via session chairs. Slides collected in advance; remote fallback is allowed via pre-recorded talks with synchronized live Q&A when possible. Standard projector/HDMI, lectern mic + 2 wireless mics for audience Q&A. Reliable Wi-Fi for live demos strongly preferred.

Recording, and Conduct. We will follow CVPR policies on recording and archival, obtain speaker consent, and encourage captioning for remote talks. All sessions adopt CVPR's Code of Conduct. We will support inclusive participation (front-row reserved seating for accessibility, mic runners for questions, and priority queueing for remote questions).

Staffing & Risk Mitigation. The team includes: two session chairs, one timekeeper, one A/V liaison, one hybrid moderator, and a posters/demos lead with two volunteers. Contingencies include: (i) slide deck centralization for fast handover; (ii) remote dial-in backup for on-site speakers; (iii) pre-recorded talks for no-shows; and (iv) a single-laptop fallback with PDF slides if live switching fails.

8. Program Plan (Half-Day Schedule)

Preliminary schedule (subject to CVPR guidelines)-

09:00–09:10	Welcome and Introduction
09:10–09:50	Keynote 1
09:50–10:30	Oral Session I (3–4 Orals, 7–8 min)
10:30–11:00	*Poster-Session & Coffee Break*
11:00–11:40	Keynote 2
11:40–12:00	Demos
12:00–13:00	Lunch

13:00–13:45 Panel: *"From Retrieval to Action: What Should Agentic Vision Systems Verify?"*
13:45–14:30 Oral Session II (3–4 Orals, 7–8 min)
14:30–15:00 Keynote 3
15:00–15:30 *Poster-Session & Coffee Break*
15:30–16:00 Keynote 4
16:30–17:00 Awards & Closing

9. Submission & Reviewing

Policy & Scope We will accept **archival, peer-reviewed** workshop papers in conjunction with CVPR 2026 proceedings, managed on **OpenReview**. Reviews are *private to the committee* (Area Chairs and assigned reviewers) and not publicly visible; comments from the general public will not be solicited. Submissions are **double-blind**; authors must have a valid OpenReview profile at submission time. Conflicts of interest (COIs) follow CVPR policy (same-employer, advisor/advisee, co-author within 3 years, or close personal relationship).

Formatting Papers should use the official CVPR style. The default limit is **8 pages (excluding references)**; appendices are allowed in the supplement only. We encourage releasing code/data/models when possible

Reviewing process Each submission receives **three expert reviews** plus **AC oversight**. A light *author response* window may be offered to clarify factual misunderstandings. Decisions will consider:

- **Technical merit & novelty** in multimodal retrieval/reranking, structured+unstructured integration, or agentic tool use.
- **Sound evaluation**: region/page/moment grounding; calibration/abstention; realistic efficiency (latency/cost/memory).
- **Reproducibility & artifacts**: completeness and clarity of released resources.
- **Broader considerations**: safety/privacy/licensing, fairness, and societal value.

Important dates (AoE)

- **CFP posted:** *Jan 6, 2026*
- **Paper submission deadline:** *Mar 5, 2026*
- **Author response (optional):** *Mar 20–22, 2026*
- **Notification to authors:** *Mar 28, 2026*
- **Camera-ready due (CVPR workshops):** *Apr 11, 2026*
- **Final program due (CVPR workshops):** *Apr 18, 2026*
- **Workshop day (Denver, USA):** *Jun 3–4, 2026*

Archival & Presentation Accepted papers will appear in the **CVPR 2026 Workshops** proceedings (subject to CVPR's camera-ready policies) and be presented as posters or short orals. At least one author must register and present in person; remote talk exceptions may be granted for documented visa or hardship cases. Non-anonymous submissions, template violations, missing license declarations for released artifacts, omission of the required Artifact Checklist, or papers lacking an ethics note when using sensitive visual data may be desk-rejected.

10. Organizers

Amit Agarwal | Oracle AI
amit.h.agarwal@oracle.com
Specializes in unstructured, structured and multilingual retrieval, and evaluation for source-linked generation. His work spans CV, IR, and NLP, with a focus on multilingual-multimodal agentic-vision systems. Amit has published at top-tier venues including ACL, NAACL, EMNLP, and ICCV. He has led cross-domain, cross-region teams and mentored underrepresented communities in AI, including Latinx and African ML groups. Amit has also served as Area Chair for EMNLP, organized workshops at ACL, and hackathons at MIT and UT Austin, and served on panels addressing AI innovation. At GRAIL-V, he serves as General Chair and Lead for Agentic-Vision Systems & Data Enrichment for Retrieval, coordinating the P→R→R→V loop baselines and cross-modality testbeds.

Vivek Gupta | Arizona State University
vgupt140@asu.edu
Assistant Professor in the School of Computing and Augmented Intelligence, specializing in NLP for structured data such as semi-structured data, flowcharts, and maps. Previously, he held roles at UPenn, University of Utah, and Microsoft Research India. His work has earned awards like the Bloomberg Data Science Fellowship and Ericsson Innovation Award. Vivek has organized the NLP for Structured Data BoF (NAACL'25, ACL'25), NAACL SRW'21, and the RARA workshop (ICDM'25). At GRAIL-V, he is General Chair and Lead for Heterogeneous Retrieval (Structured Sources), curating structured/GUI/doc tasks and evaluation protocols for parsing tools, retrieval and reranking.

Vivek Srikumar | University of Utah
svivek@cs.utah.edu
Associate Professor at the Kahlert School of Computing, co-leads Utah NLP and is affiliated with the Utah Center for Data Science. His research spans machine learning and NLP, with support from NSF, NIH, and industry (Google, Intel, NVIDIA). He has held roles at AI2, Stanford NLP, and UIUC. Vivek has served as Program Co-Chair for ACL 2024 and CoNLL 2022, and has organized workshops at NeurIPS, EMNLP, and AAAI. He is currently an Action Editor at Computational Linguistics and a member of the JAIR board. At GRAIL-V, he co-leads for Grounding & Reliability, defining the review rubric for grounded verification, calibration/abstention, and end-to-end reliability.

Tao Sheng | Oracle AI
tao.t.sheng@oracle.com
Researcher and technologist with 15+ years of industrial R&D experience spanning Computer Vision, GenAI, LLMs, RAG, and AI agents and Analytics. He has held leadership roles at organizations including Oracle, Amazon, Qualcomm, and Intel, contributing to the development of large-scale AI products. Community leadership includes roles across CVPR, ACL, ICCV, ECCV and NeurIPS. Tao holds 45+ patents and has authored over a dozen peer-reviewed research papers, actively contributing to the advancement of the CV field. At GRAIL-V he leads Agentic-Vision Planning & Tool Use.

Alice Oh | KAIST Computer Science
alice.oh@kaist.edu
KAIST ICT Chair Professor in the School of Computing at KAIST, with a joint appointment in the Graduate School of AI. Her research lies at the intersection of NLP and computational social science, with a recent focus on multilingual and multicultural aspects of LLMs. She collaborates with scholars in political science, education, and history. Alice has served as Program Chair for ICLR 2021 and NeurIPS 2022, General Chair for ACM FAccT 2022 and NeurIPS 2023, and DEI Chair for COLM 2024. She is the current President of SIGDAT, overseeing EMNLP. At GRAIL-V, she co-leads the Attribution & Multilingual tracks, ensuring multilingual coverage, attribution standards, and DEI-aligned evaluation practices across multimodal methods.

Sara Hooker | Co-Founder Adaption Labs, ex-VP Research Cohere Labs
sarahookr@gmail.com
Sara Hooker is a co-founder of Adaption Labs, which builds intelligence that continuously evolves. Sara leads a large team of AI researchers and engineers that build extremely efficient, adaptable systems. Sara Hooker was previously VP of Research at Cohere, a $6.8 billion frontier AI company focused on generative AI for enterprise. Prior to Cohere, she built large systems in Computer Vision and NLP at Google Deepmind. Her research has been published in top venues including Nature, NeurIPS, ICML, ACL, ICLR, EMNLP, MLSys and has been recognized with honors such as the ACL Best Paper Award and CACM front cover for her work on the Hardware Lottery. Her work has been featured in mainstream news outlets including Techcrunch, New York Times, Washington Post, Axios, MIT Technology, The Atlantic. Sara is a frequent expert advisor to AI research and policy initiatives around the world: she is currently on Kaggle's ML Advisory Research Board and serves on the World Economic Forum council on the Future of Artificial Intelligence and the Future of Data Frontiers. She has been listed as one of AI's top 13 innovators by Fortune and one of Time100 Most Influential People in AI. At GRAIL-V, she chairs Agentic-Vision Systems & Efficiency and Industry Liaison, driving deployment-aware evaluation (latency/memory/cost/energy) and industry demos aligned with the P→R→R→V loop.

Jyotika Singh | Oracle AI
jyotika.s.singh@oracle.com
Specializes in NLP, conversational AI agents, and text-to-SQL systems. Jyotika is an author of an NLP book (CRC Press) with several patents in applied AI. A frequent member of international conference committees (ACL, EMNLP, NAACL, and others), speaker at global AI venues, and adjunct assistant professor at University of Southern California, she is also an active mentor for women and underrepresented groups in AI. At GRAIL-V, she co-leads Agentic-Vision Systems for long-term/short-term memory and human-interaction, operationalizing agents/pipelines and showcasing end-to-end agentic evaluations.

Hitesh Patel | Oracle AI
hitesh.laxmichand.patel@oracle.com
Hitesh specializes in guardrails across multilingual NLP, multimodal language models, and information retrieval. He collaborates with academic labs and industry partners such as CILVR @ NYU, SeaCrowd, MBZUAI, Cohere Labs, and Data Intelligence Lab. Hitesh has organized workshops and contributed extensively to the multilingual and multimodal NLP communities. At GRAIL-V, he co-leads the Multilingual & Multimodal track with a focus on responsible AI and

guardrail evaluations (safety, leakage, abstention) for real-world deployments.

11. Program Committee

The seed list (expanding to ∼60–75) ensures 3 reviews per paper with balanced loads; organizers will add domain experts as needed and avoid COIs.

Members (seed list; to be further expanded):

(1) **Srikant Panda** (Optum) — Agentic AI, Responsible AI; (2) **Hitesh Patel** (Oracle) — Multimodal Responsible AI & IR (3) **Karan Dua** (Oracle) — Benchmarking and Evaluation ;(4) **Olena Burda-Lassen** (Sonepar; Taras Shevchenko Univ. of Kyiv) — Multimodal & Responsible AI ; (5) **Meizhu Liu** (Oracle; University of Florida) — Unstructured Retrieval ; (6) **Ziyan Jiang** (UC Santa Barbara) — Multimodal IR ; (7) **Brian Lin** (Instacart; UC Berkeley) — Applied NLP and IR ; (8) **Gauri Kholkar** (Pure Storage; BITS) — Applied NLP & IR ; (9) **Marcela Medicina Ferreira** (Univ. Estadual de Campinas) — AI in Medicine ; (10) **Michael Avendi** (Oracle; UC Irvine) — Multimodal LLMs ; (11) **Yassi Abbasi** (Oracle; USC) — IR (12) **Haodong Duan** (Shanghai AI Lab) — Multimodal AI ; (13) **Ulrich Bodenhofer** (QUOMATIC.AI; Univ. of Applied Sciences, Austria) — ML/DL, Medical/Bioinformatics ; (14) **Hasan Iqbal** (MBZUAI) — NLP/IR ; (15) **Jaewon Jung** (Seoul National University) — Robustness in LLMs ; (16) **Nasib Ullah** (Aalto University) — IR, Multimodal ; (17) **Kunal D** (NVIDIA) — Multimodal AI ; (18) **Hank Lee** (CMU) — Agentic AI ; (19) **Neil Shah** (Snapchat) — Data mining . (20) **Praneet Pabolu** (Splunk) — Agentic AI ; (21) **Nirmesh Shah** (Sony Research) — Multimodal speech ; (22) **Nishanth Madhusudhan** (ServiceNow) — LLM eval & benchmarking (23) **Osman Alperen Koraş** (Inst. for AI in Medicine) — Agentic AI for medicine ; (24) **Akhil Arora** (Aarhus University) — Agentic AI ;(26) **Koustuv Saha** (University of Illinois) — Ethics in AI .

12. Diversity and Inclusion

We commit to a workshop that is broadly accessible and representative across demographics, geography, seniority, and institution type, and that aligns with CVPR 2026 guidance on broad representation, ethical considerations, and hybrid participation.

Organizers and Speakers. Our Organizing Committee and invited slate are balanced across gender, seniority (student → senior leader), and affiliation (academia/industry/non-profit), with meaningful participation from the Global South. We will coordinate with the CVPR workshop track structure to minimize overlap and ensure a unique lineup, in line with CVPR's request to avoid speaker duplication across workshops.

Program Committee and Reviewing. We will recruit a geographically distributed PC with a mix of first-time and experienced reviewers, provide a short inclusive-reviewing briefing (rubric + examples for low-resource and non-Western contexts), enforce COIs, and balance loads. If we include archival proceedings, we will follow CVPR's workshop paper requirements (style/length/timeline) and provide limited shepherding to support authors new to CVPR.

Content Scope. We explicitly encourage work in under-represented settings (low-resource/sign languages, non-Latin scripts, domain shift, and non-Western datasets), transparent reporting (model cards/datasheets when applicable), and the release of artifacts (code, data, slides, eval scripts) under permissive licenses.

Participation and Outreach. CFP and reviewer calls will be distributed via WiCV, Black in AI, LatinX in AI, Queer in AI, WiML, Deep Learning Indaba, and Masakhane. We will run light-touch clinics (pre-submission abstract checks; camera-ready polishing) prioritizing first-time authors and under-represented regions.

Accessibility and Hybrid. In accordance with CVPR 2026, the event will be in-person with hybrid support (Zoom room provided by CVPR). We will offer moderated virtual Q&A, pre-recorded talk options, and asynchronous discussion channels to include participants across time zones; we will request captions and share slide-accessibility guidance (readable fonts/contrast; alt text for key figures). We acknowledge CVPR's policy that fully virtual workshops are not permitted.

Accountability. Post-workshop, we will publish an anonymized summary of participation statistics (submissions, acceptances, speaker/PC demographics in aggregate, geographic spread) and lessons learned to inform future editions.

13. Ethical Considerations

GRAIL-V enforces privacy-preserving use of visual data and grounded, reproducible reporting. Submissions must favor redaction or synthetic surrogates for PII/PHI and use controlled-access evaluation for sensitive screenshots/forms/logs; region/moment-level evidence overlays are required to avoid revealing unre-

lated content. To support fairness and inclusion, authors are encouraged to use multilingual, layout-diverse datasets and report failure modes—including calibration/abstention behavior—across domains and visual styles. Because agentic systems may invoke external tools, papers must document safeguards (e.g., sandboxing, allow-lists, rate limits) and assess prompt-injection and leakage risks specific to UIs/DOM and copy–paste channels. All artifacts must state provenance, usage rights, and license compatibility (datasets, code, models; screenshots/web content with ToS/robots policies and transformation rationale).

For responsible release and review, each submission includes an artifact checklist: (i) licenses per artifact; (ii) evidence overlays/spans/moments supporting claims; (iii) calibration/abstention metrics; (iv) efficiency numbers (latency, memory, cost/energy); and (v) a summary of safety tests (e.g., leakage/prompt-injection). Submissions involving sensitive data must include an ethics statement; reviewers may flag risk, triggering an organizer ethics review. The workshop follows a strict conflict-of-interest policy for organizers, speakers, and the program committee.

References

[1] O-drum @ cvpr 2022: Workshop on open-domain retrieval under multi-modal settings. https://asu-apg.github.io/odrum/archive_2022.html, 2022. Workshop website. 3

[2] Mula 2023: 6th multimodal learning and applications workshop (in conjunction with cvpr 2023). https://mula-workshop.github.io/index_2023.html, 2023. Workshop website. 3

[3] Tavi@cvpr'24: Tool-augmented vision workshop. https://sites.google.com/view/tavi-cvpr24/, 2024. Workshop website. 3

[4] Wild3d: 3d modeling, reconstruction, and generation in the wild (eccv 2024 workshop). https://3d-in-the-wild.github.io/2024.html, 2024. Workshop website. 3

[5] Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, and Rainer Stiefelhagen. Rodla: Benchmarking the robustness of document layout analysis models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15556–15566, 2024. 1

[6] DriveX Organizing Committee. Drivex: Foundation models for v2x-based cooperative autonomous driving (2nd edition). https://drivex-workshop.github.io/, 2025. ICCV 2025 Workshop. Accessed Oct 30, 2025. 3

[7] Kuicai Dong, Yujing Chang, Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multi-modal retrieval for long documents, 2025. 1

[8] IEEE/CVF. Cvpr 2026 call for workshops. https://cvpr.thecvf.com/Conferences/2026/CallForWorkshops, 2025. Accessed Oct 30, 2025. 1, 3

[9] Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. Genius: A generative framework for universal multimodal search. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19659–19669, 2025. 1

[10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022. 1

[11] Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. Enhancing visual document understanding with contrastive learning in large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15546–15555, 2024. 1

[12] Yichen Li, Congyue Deng, Katie Luo, Yonglong Tian, et al. Geometry in large model era (geolme). https://geo-lme.github.io/, 2024. ECCV 2024 Workshop. Accessed Oct 30, 2025. 3

[13] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025, 2025. 1

[14] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15630–15640, 2024. 1

[15] Wei-Chiu Ma, Shenlong Wang, Yufei Ye, Linyi Jin, et al. Wild3d: 3d modeling, reconstruction, and generation in the wild. https://3d-in-the-wild.github.io/, 2025. ICCV 2025 Workshop. Accessed Oct 30, 2025. 3

[16] Quentin Macé, António Loison, and Manuel Faysse. Vidore benchmark v2: Raising the bar for visual retrieval, 2025. 1

[17] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19446–19455, 2025. 1

[18] Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering.

In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria, 2025. Association for Computational Linguistics. 1

[19] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024. 1

[20] Zilin Xiao, Pavel Suma, Ayush Sachdeva, Hao-Jen Wang, Giorgos Kordopatis-Zilos, Giorgos Tolias, and Vicente Ordonez. Locore: Image re-ranking with long-context sequence modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9580–9590, 2025. 1

[21] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15116–15127, 2023. 1