Thematic session

Paper presentation: Group1

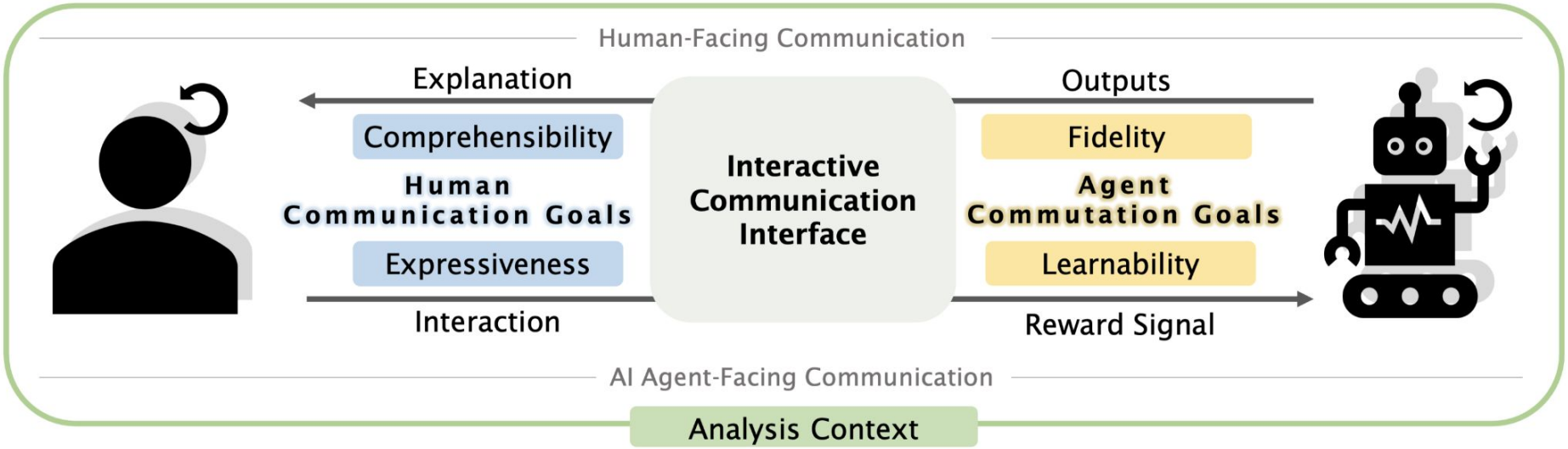
Marcel Worring

Mapping out the Space of Human Feedback for Reinforcement Learning: A Conceptual Framework

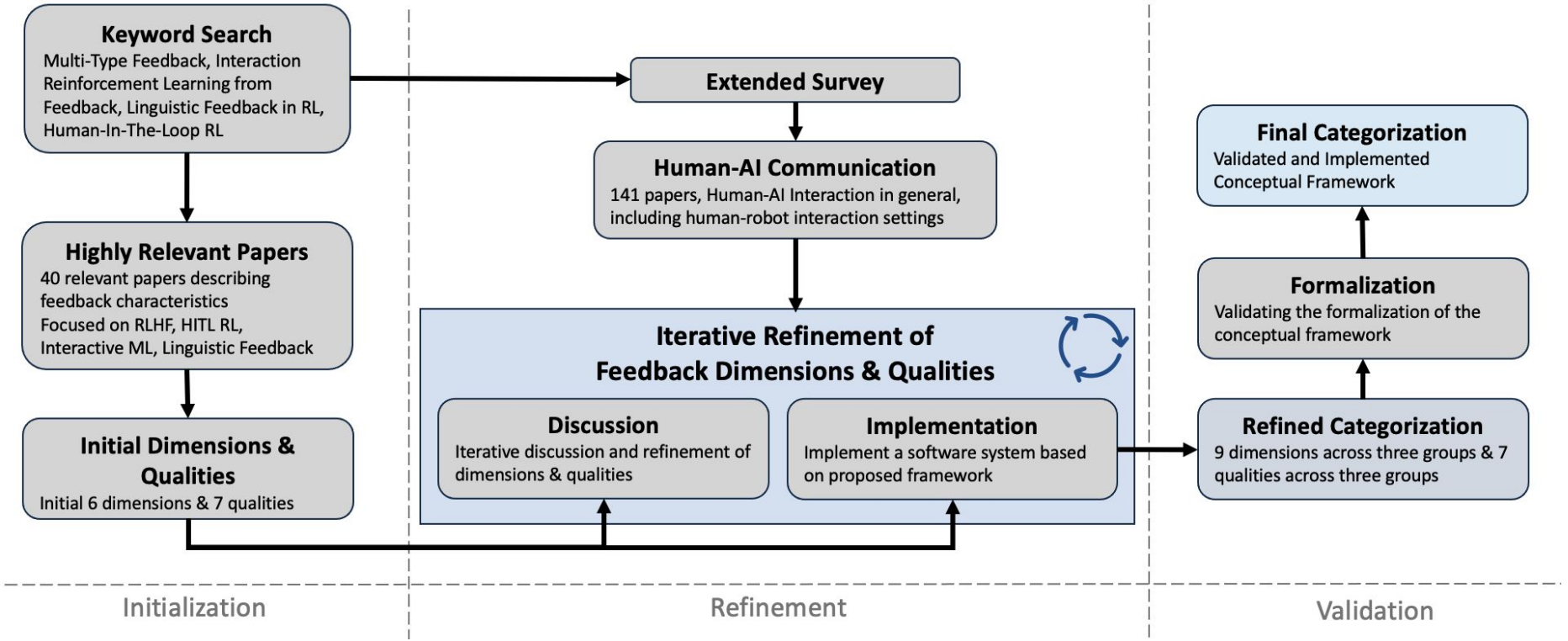
YANNICK METZ, University of Konstanz, Germany and ETH Zurich, Switzerland
DAVID LINDNER, ETH Zurich, Switzerland
RAPHAËL BAUR, ETH Zurich, Switzerland
MENNATALLAH EL-ASSADY, ETH Zurich, Switzerland

arXiv:2411.11761v2 [cs.LG] 20 Feb 2025

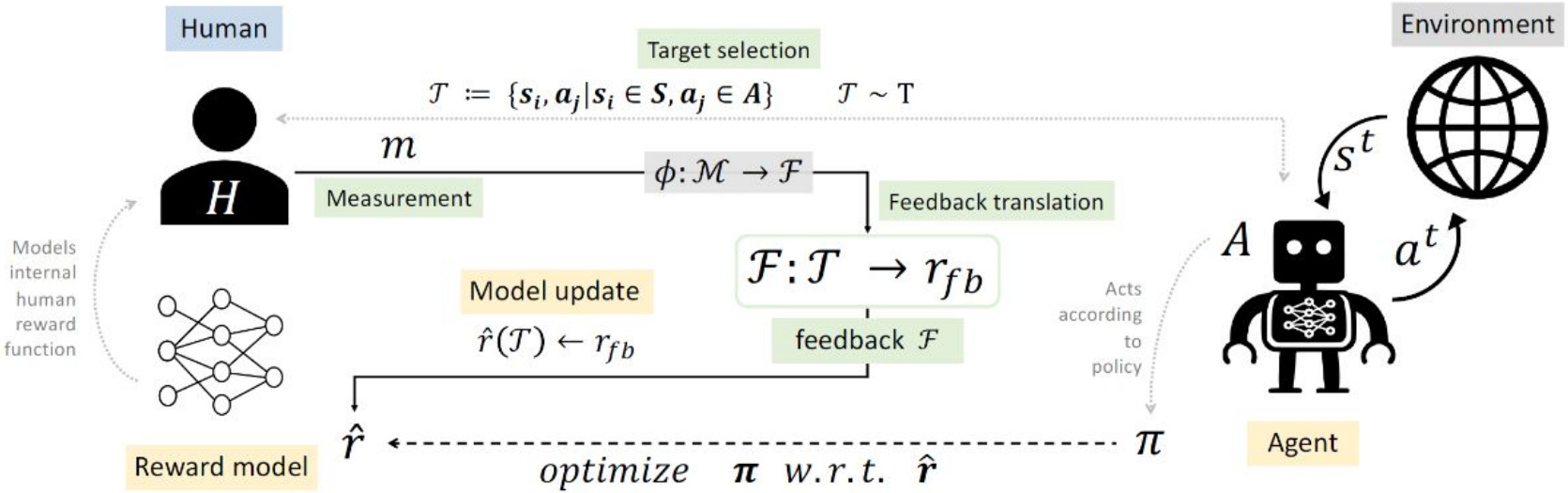
Human-AI Interaction



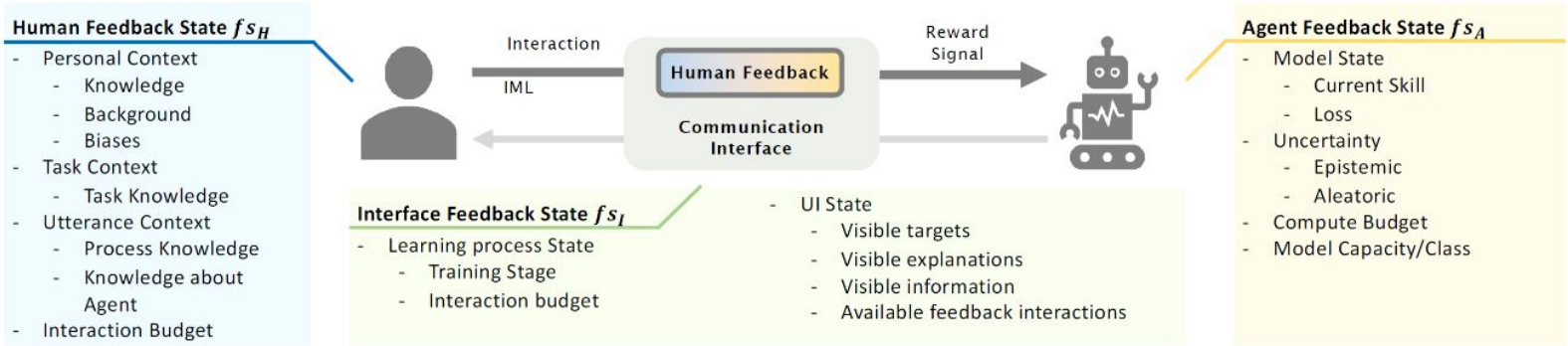
Creating a Review



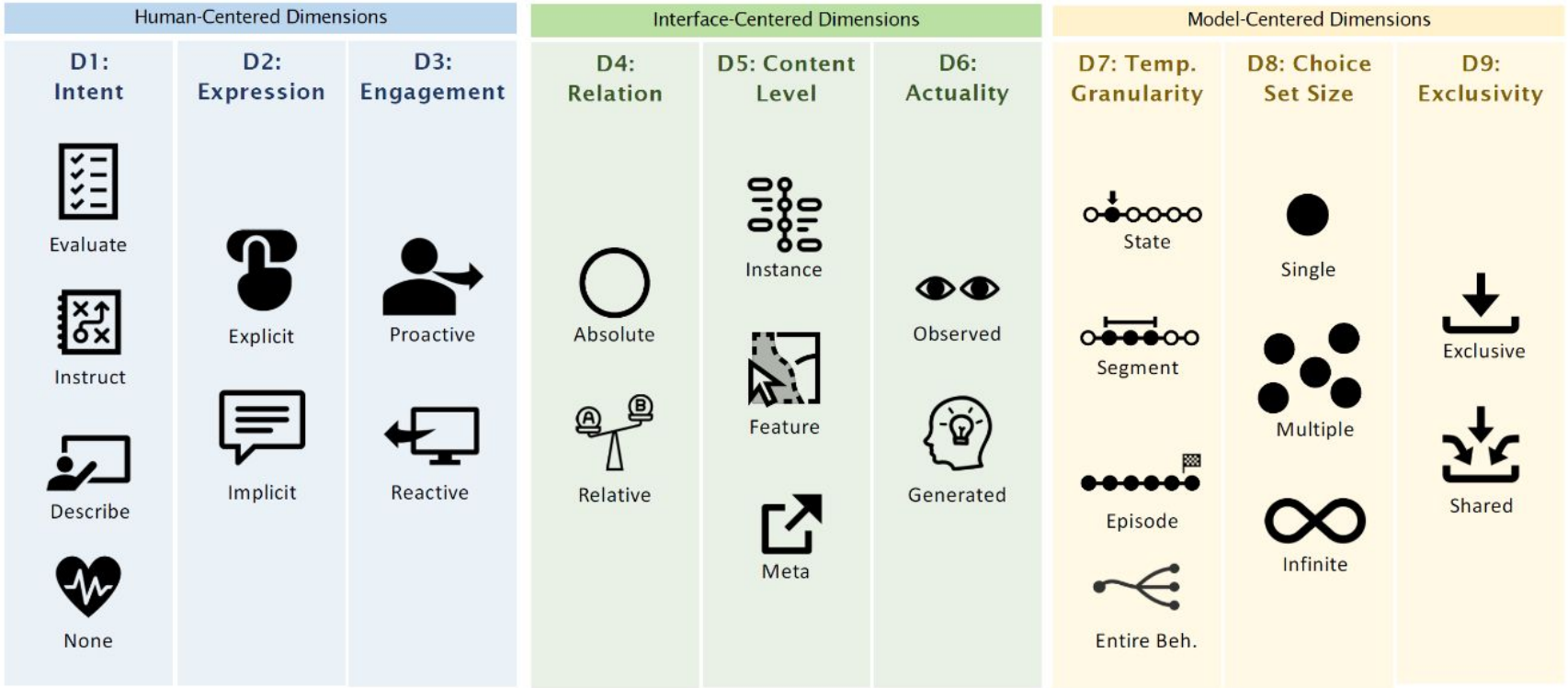
A more formal model



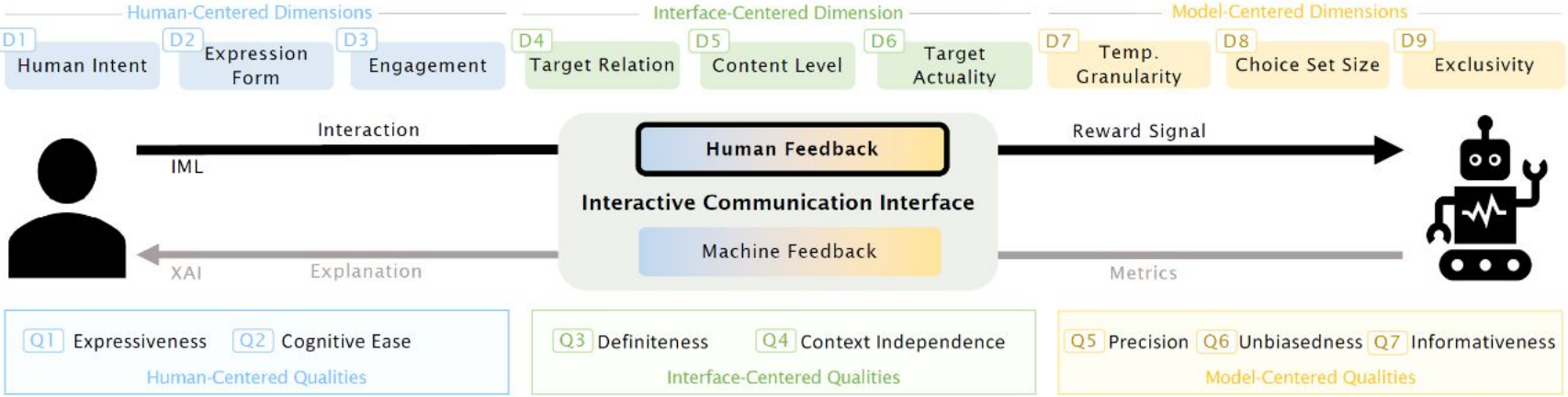
Different State Spaces



Feedback Dimensions



Summary


Human-Centered Dimensions
Interface-Centered Dimension
Model-Centered Dimensions
D1 Human Intent
D2 Expression Form
D3 Engagement
D4 Target Relation
D5 Content Level
D6 Target Actuality
D7 Temp. Granularity
D8 Choice Set Size
D9 Exclusivity
Interaction
IML
Human Feedback
Interactive Communication Interface
Machine Feedback
Reward Signal
XAI
Explanation
Metrics
Q1 Expressiveness
Q2 Cognitive Ease
Human-Centered Qualities
Q3 Definiteness
Q4 Context Independence
Interface-Centered Qualities
Q5 Precision
Q6 Unbiasedness
Q7 Informativeness
Model-Centered Qualities

Shuai Wang

Published as a conference paper at ICLR 2025

ColPali: Efficient Document Retrieval with Vision Language Models

Manuel Faysse[*1,3] **Hugues Sibille**[*1,4] **Tony Wu**[*1] **Bilel Omrani**[1]
Gautier Viaud[1] **Céline Hudelot**[3] **Pierre Colombo**[2,3]
[1]Illuin Technology [2]Equall.ai [3]CentraleSupélec, Paris-Saclay [4]ETH Zürich
manuel.faysse@centralesupelec.fr

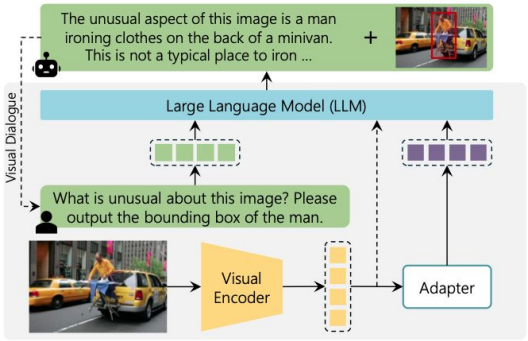
Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., & Colombo, P. (2024). ColPali: Efficient Document Retrieval with Vision Language Models. *ArXiv, abs/2407.01449*.

Background –Multimodal Large Language Models (MLLMs) vs. Large Language Models (LLMs)

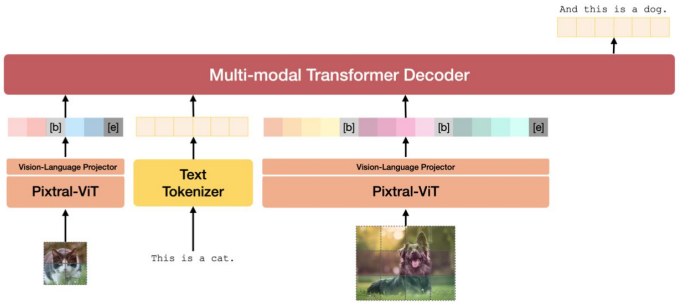
MLLMs are a **superset** of **Large Language Models LLMs**:

✅ They have the **added ability** to process and understand **images, charts, tables, and figures**.

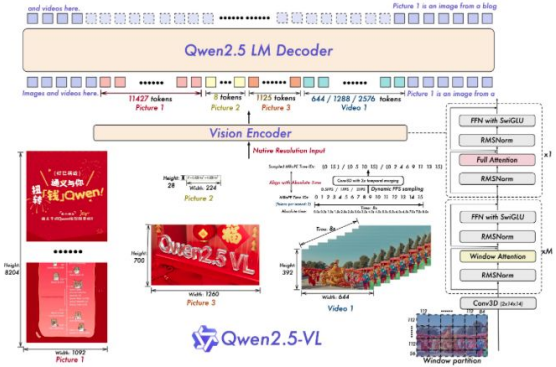
✅ They typically use an **LLM decoder as the backbone**, should retain **all text-based capabilities** (includes reasoning) of LLMs.



General architecture of MLLM



Pixtral 12B from 2024 Oct



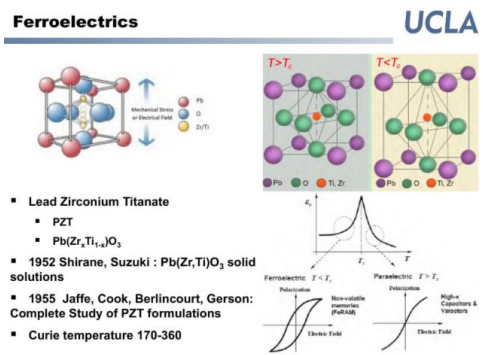
Qwen2.5 VL from 2025 Jan

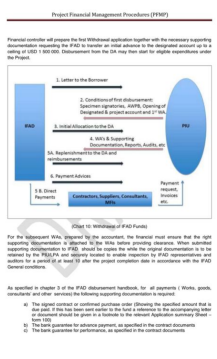
Why MLLMs for Document Understanding?

📄 **Documents are inherently multimodal**—they combine **text, charts, tables, and figures**, where LLM struggle to extract insights from.

🌍 **Documents represent knowledge across different domains with diverse formats** – Scientific papers, financial reports, legal contracts, and medical records…

🚀 **MLLMs unlock a new frontier** in **document retrieval, summarization, and comprehension** by handling all these modalities **simultaneously**.

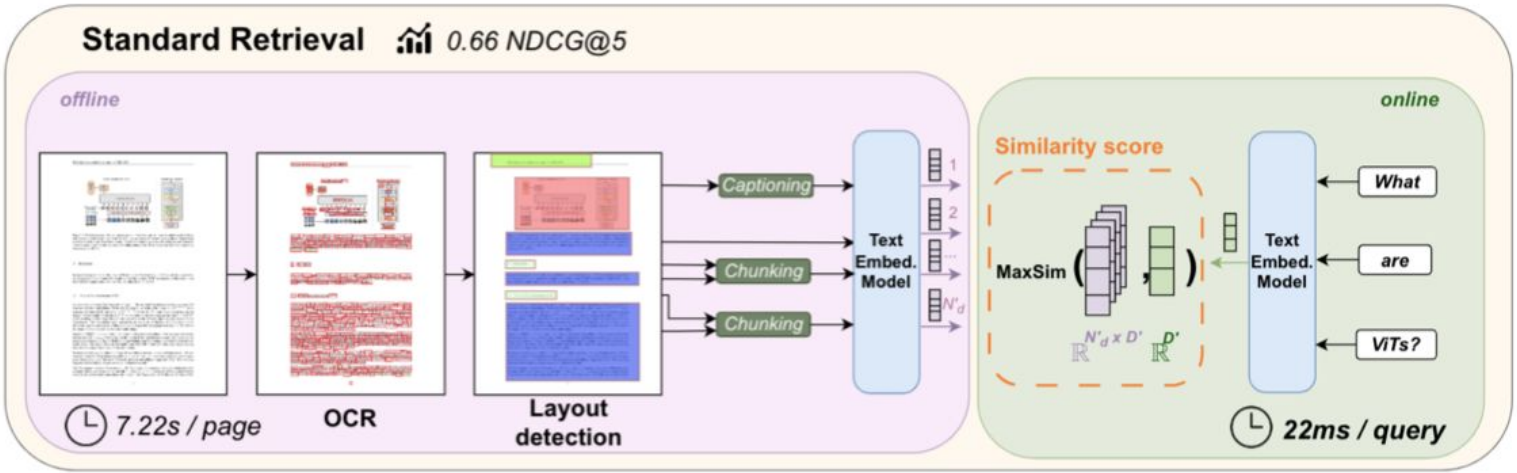




Scientific papers, slides, reports…

Main Paper Idea:

- Current document retrieval methods rely heavily on text extraction (OCR, parsing), **neglecting visual cues**.
- **ColPali** proposes a **vision-based retrieval model** that indexes document pages **directly from images** using **Vision-Language Models (VLMs)**.
- **Outcome:** ColPali is **faster, simpler, and more accurate** than conventional retrieval systems.



How ColPali Works?

Multi-Vector Vision Embeddings:

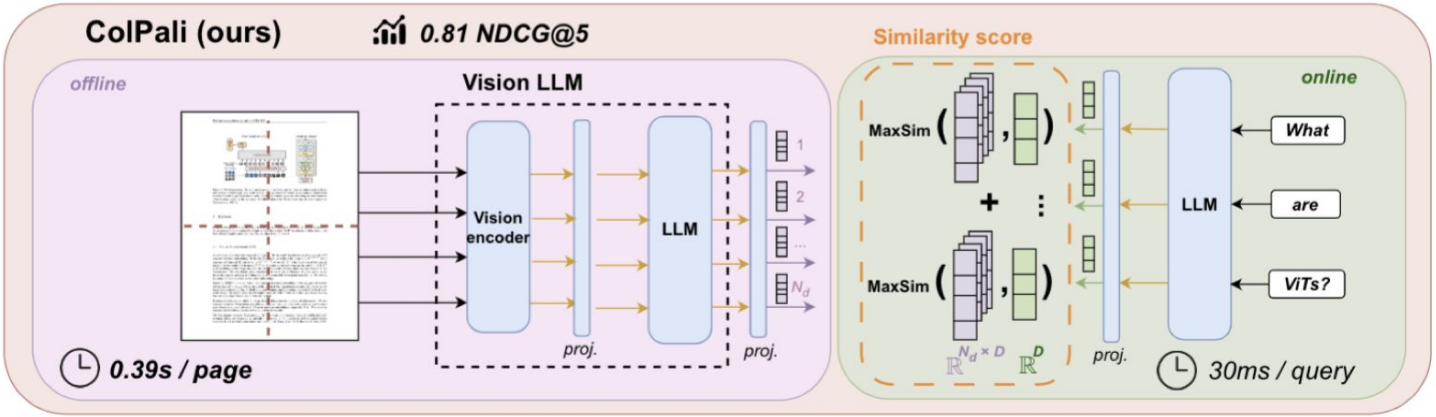
- Uses **PaliGemma-3B**, a Vision-Language Model, to encode images and query into **multi-vector embeddings**

Late Interaction Mechanism:

- Inspired by **ColBERT**, performs **fine-grained** matching between query and document embeddings.

End-to-End Training with Contrastive Learning:

- **Trained on 118K query-page pairs** (academic + synthetic datasets).



ViDoRe: A New Benchmark for Document Retrieval

Introduced in this paper to evaluate retrieval performance across:

- Different **document types** (academic, administrative, scientific).
- **Multiple modalities** (text, tables, infographics, figures).
- **Various languages** (English, French).
- **Two evaluation categories:**
  - **Academic Tasks:** Repurposed **VQA datasets** (e.g., DocVQA, InfoVQA, arXivQA).
  - **Practical Tasks:** Domain-specific retrieval benchmarks (e.g., government, healthcare, energy).

Dataset	Language	# Queries	# Documents	Description
Academic Tasks				
DocVQA	English	500	500	Scanned documents from UCSF Industry
InfoVQA	English	500	500	Infographics scrapped from the web
TAT-DQA	English	1600	1600	High-quality financial reports
arXiVQA	English	500	500	Scientific Figures from arXiv
TabFQuAD	French	210	210	Tables scrapped from the web
Practical Tasks				
Energy	English	100	1000	Documents about energy
Government	English	100	1000	Administrative documents
Healthcare	English	100	1000	Medical documents
AI	English	100	1000	Scientific documents related to AI
Shift Project	French	100	1000	Environmental reports

Table 1: *ViDoRe* comprehensively evaluates multimodal retrieval methods.

5 RESULTS

	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured text-only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4↓5.7	-	-	36.1↓7.9	68.5↑8.9	88.4↓2.0	76.8↓1.5	77.7↓1.1	84.6↑2.0	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4↓0.2	25.7↓11.1	60.1↓2.8	70.8↑24.3	50.5↓12.2	**73.2**↑8.9	90.2↓2.6	83.6↓2.3	84.9↑1.0	91.1↑3.9	66.1↑0.6
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7↓4.4	32.9↓5.4	71.9↑1.9	69.1↑33.7	43.8↓17.7	73.1↑12.2	88.8↑0.8	83.3↓1.4	80.4↓2.3	91.3↑2.1	67.0↑1.9
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	58.5↑15.3	32.9↑2.6	70.5↑6.4	62.7↑4.6	30.5↑4.3	26.5↑7.8	74.3↑11.8	73.7↑8.0	74.2↑8.1	82.3↑3.2	58.6↑7.2
BiPali (+LLM)	56.5↓-2.0	30.0↓-2.9	67.4↓-3.1	76.9↑14.2	33.4↑2.9	43.7↑17.2	71.2↓-3.1	61.9↓-11.7	73.8↓-0.4	73.6↓-8.8	58.8↑0.2
ColPali (+Late Inter.)	**79.1**↑22.6	**54.4**↑24.5	**81.8**↑14.4	**83.9**↑7.0	**65.8**↑32.4	**73.2**↑29.5	**96.2**↑25.0	**91.0**↑29.1	**92.7**↑18.9	**94.4**↑20.8	**81.3**↑22.5

Table 2: **Comprehensive evaluation of baseline models and our proposed method on *ViDoRe*.** Results are presented using nDCG@5 metrics, and illustrate the impact of different components. Text-only metrics are not computed for benchmarks with only visual elements.

Take away message:

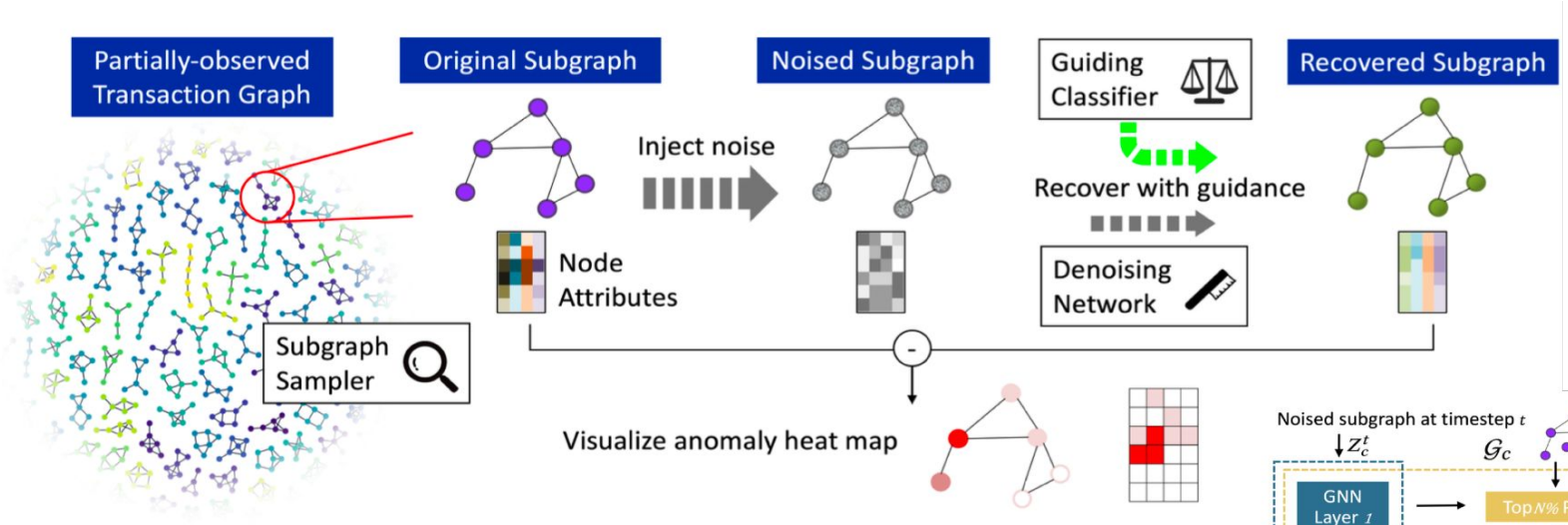
- ColPali **simplifies and enhances document retrieval** by leveraging Vision-Language Models (VLMs) t, eliminating the need for text extraction.

- ColPali also shows a **way to MLLM application into different specific domains**. This could go beyond retrieval to **knowledge extraction**, **Forensic/Financial/Cultural industry document analysis**, and **AI-powered research assistants**.

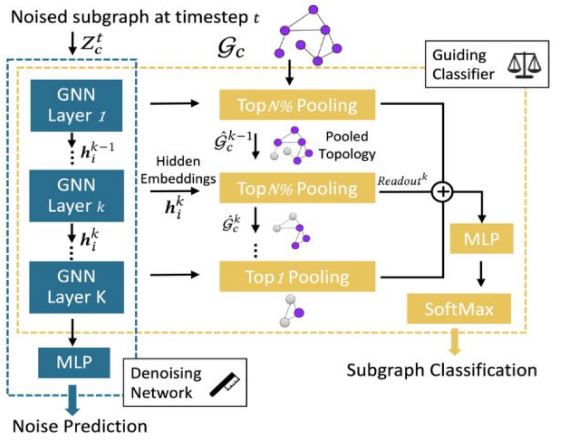
Yijia Zheng

Yassin Mohamadi

Diga: Guided Diffusion Model for Graph Recovery in Anti-Money Laundering



KDD 2023



Li, Xujia, et al. "Diga: Guided diffusion model for graph recovery in anti-money laundering." *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023.

DATASETS

Dataset	WB-S	WB-M	WB-L	Elliptic	OTC	Alpha
# Nodes	58,409	233,638	467,277	203,769	5,858	3,754
# Edges	101,910	523,301	904,256	234,355	35,592	24,186
# Node Features	22	22	22	166	4	4
# Anomalies	478	1,830	3,676	4,545	178	102
Anomaly Ratio (%)	0.81	0.78	0.79	2.23	3.03	2.72
Avg. Degree	1.74	2.24	1.93	1.15	6.06	6.44

Results

Dataset	WeBank-small		WeBank-medium		WeBank-large		Elliptic		OTC		Alpha	
Metrices	Pre@100	AUC(%)	Pre@100	AUC(%)	Pre@100	AUC(%)	Pre@100	AUC(%)	Pre@100	AUC(%)	Pre@100	AUC(%)
SVM	0.162 ±0.021	54.0 ±1.5	-	-	-	-	-	-	0.519 ±0.002	68.3 ±0.1	0.477 ±0.001	63.8 ±0.3
XGBoost 2016	0.281 ±0.008	62.7 ±0.2	0.432 ±0.011	61.4 ±0.5	-	-	0.813 ±0.001	81.8 ±0.3	0.678 ±0.000	85.9 ±0.0	0.606 ±0.000	81.2 ±0.0
DeepFD 2018	0.126 ±0.037	49.2 ±2.1	0.666 ±0.038	69.4 ±3.6	0.613 ±0.049	65.7 ±4.6	0.688 ±0.033	72.3 ±3.9	0.656 ±0.031	74.6 ±2.2	0.369 ±0.022	54.7 ±3.4
SemiADC 2021	0.170 ±0.017	56.3 ±0.9	0.717 ±0.020	76.9 ±1.2	-	-	-	-	0.545 ±0.051	75.2 ±1.8	0.522 ±0.029	74.1 ±0.9
OCGTL 2022	0.267 ±0.006	60.3 ±0.1	0.802 ±0.009	77.7 ±0.1	0.751 ±0.012	73.4 ±0.4	0.912 ±0.007	90.3 ±1.1	0.622 ±0.001	83.6 ±0.6	0.683 ±0.005	80.7 ±0.1
ComGA 2022	0.331 ±0.020	71.9 ±0.6	0.804 ±0.016	80.8 ±2.0	-	-	0.855 ±0.029	85.2 ±3.6	0.701 ±0.005	87.0 ±2.4	0.634 ±0.016	84.1 ±1.9
Diga (Ours)	**0.383** ±0.004	**74.3** ±0.3	**0.900** ±0.010	**87.6** ±0.2	**0.949** ±0.008	**82.5** ±0.4	**0.981** ±0.005	**95.2** ±0.7	**0.730** ±0.010	**89.1** ±0.3	**0.724** ±0.019	**87.4** ±0.5

Gonçalo Marcelino

Understanding Barriers to Network Exploration with Visualization: A Report from the Trenches

Mashael AlKadi, Vanessa Serrano, James Scott-Brown, Catherine Plaisant,
Jean-Daniel Fekete, Uta Hinrichs and Benjamin Bach

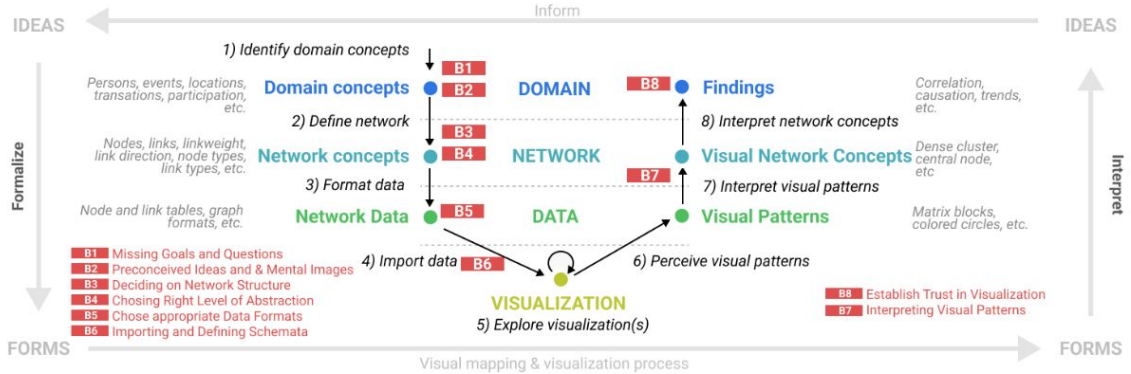
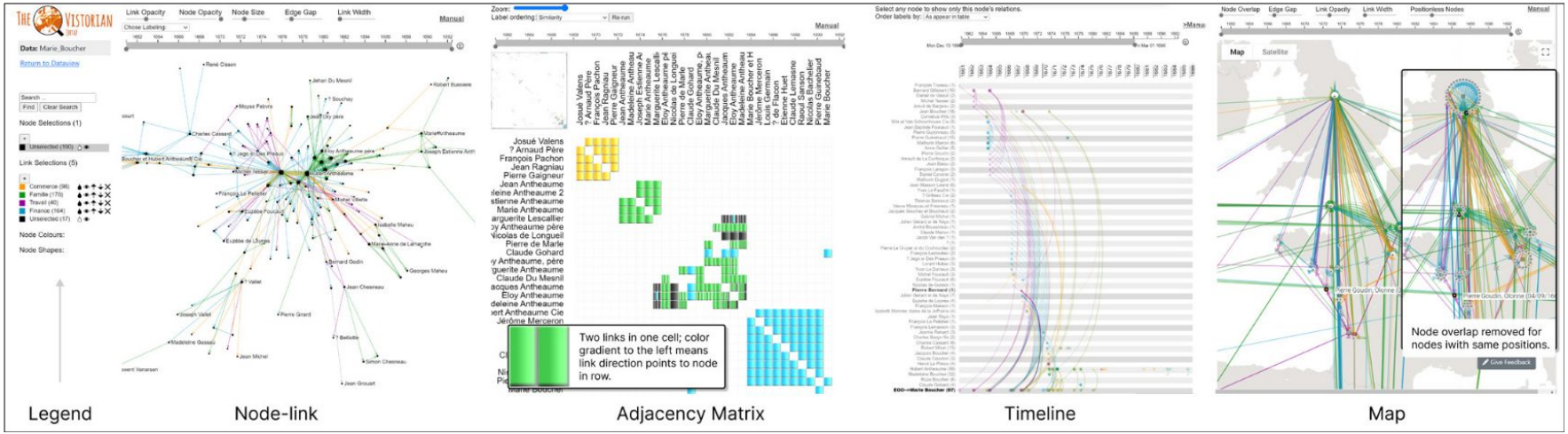


Fig. 1: Barriers (red) that may be encountered during visual network exploration process, while translating domain concepts (ideas) into network structures and visualizations (forms) and back into findings.

Abstract—This article reports on an in-depth study that investigates barriers to network exploration with visualizations. Network visualization tools are becoming increasingly popular, but little is known about how analysts plan and engage in the visual exploration

Network visualization tools are becoming increasingly popular. How do users engage in the visual exploration of network data, which exploration strategies they employ, and how they prepare their data, define questions, and decide on visual mappings?

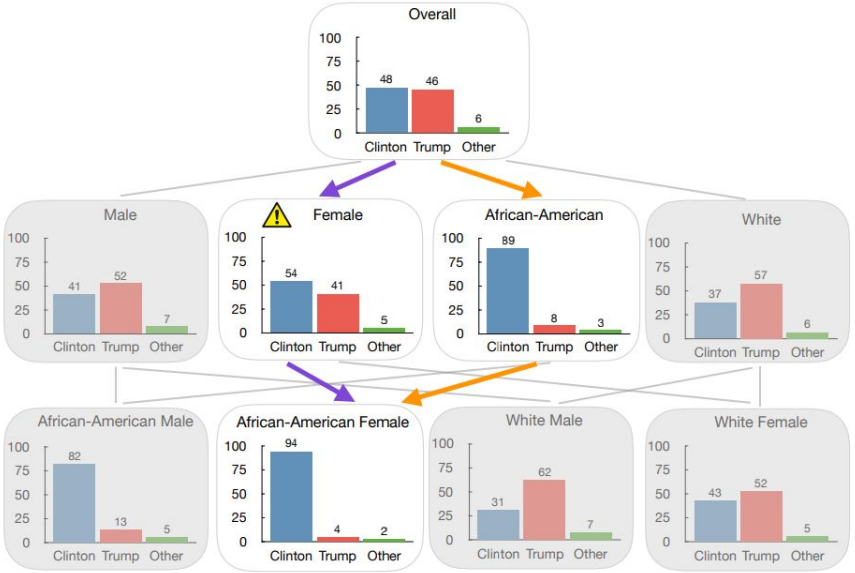


Study 1 - Researchers tracked users of the Vistorian logging 534 sessions to understand how it is being used.

Study 2 - Researchers collected qualitative during during a 6-week network exploration course by monitoring 36 participants. 50% without experience in network visualization.

- Missing Goals & Questions
  - ✗ Choosing schemas and visualizations is difficult without specific goals and can lead to drill down fallacy.
  - ✓ Sketching, and examples improved results.

- Preconceived Ideas & Mental Images / Deciding on a Network Structure
  - ✗ Wrong preconceived about what a graph visualization is (e.g. exclusively a social network graph)
  - ✓ Guided process to construct schemas and fast sketching.



Lee, Doris Jung-Lin, et al. "Avoiding drill-down fallacies with vispilot: Assisted exploration of data subsets." *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 2019.

- Choosing The Right Level of Abstraction
  - ✗ Choosing the wrong level of abstraction can lead to cluttered graphs.
  - ✓ Transformation and aggregation strategies.
- Interpreting Visual Patterns in Visualization
  - ✗ Understanding patterns in data is complex specially when interaction can lead to changes of visual patterns on-the-fly.
  - ✓ Support multiple coordinated views, support for examples.
- Establish Trust in a Network Visualization
  - ✗ Unfamiliar visualization (e.g. adjacency matrix) and misunderstanding provenance.
  - ✓ Showing examples of use in credible sources (e.g.) journalism, explaining algorithms, explaining anti-patterns.

Fatemeh Gholamzadeh