Attention Grounded Enhancement for Visual Document Retrieval

Wanqing Cui[1], Wei Huang[2], Yazhi Guo[1], Yibo Hu[1],
Meiguang Jin[1], Junfeng Ma[1], Keping Bi[2]

1. Alibaba Group, Beijing, China
2. University of Chinese Academy of Sciences, Beijing, China

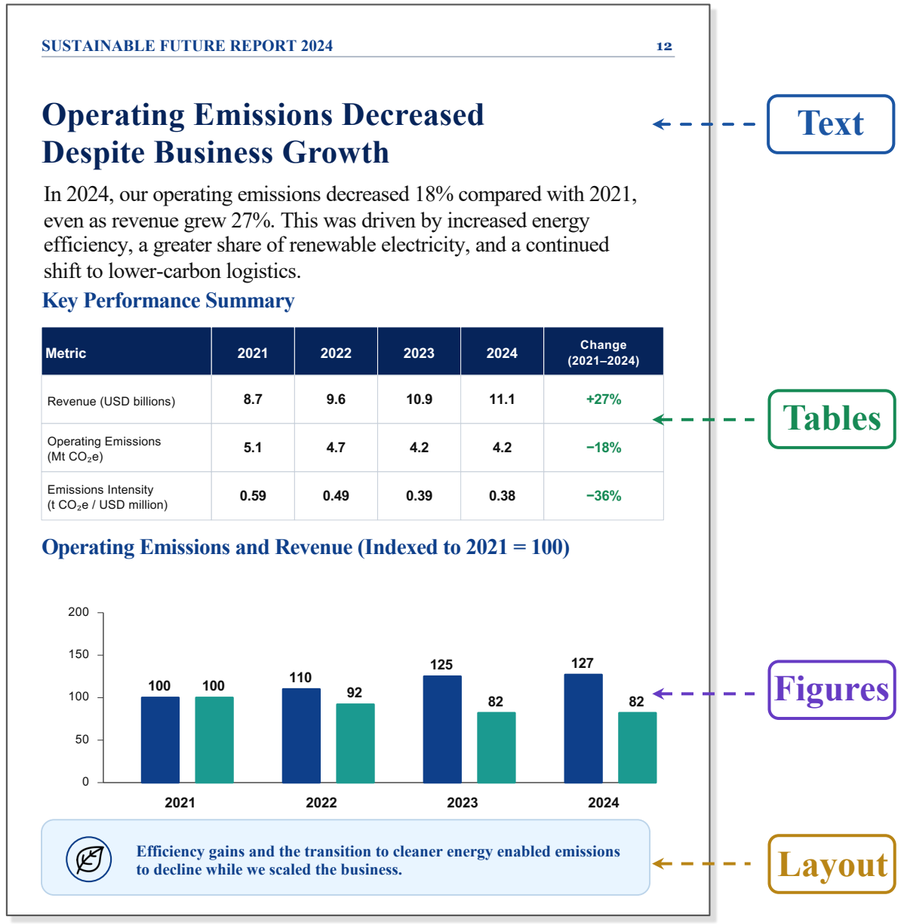
Contents

Background: Visual Document Retrieval Requires Page-Level Understanding

QUERY

Which report explains why operating emissions decreased despite business growth?

Natural-language information need



Why is this hard?

- The explanation may be implicit
- Evidence spans regions
- Layout carries meaning

The retriever must understand the page, not merely read its text.

Current Paradigm: Fine-Grained Late Interaction

Natural-language query

What does the symbol ‘ε’ represent in the bottleneck section of the depicted autoencoder architecture?

Query-token embeddings

autoencoder

epsilon

bottleneck

Token–Patch Similarity + MaxSim

autoencoder	0.18	0.27	0.82	0.25	0.41	0.12
epsilon	0.05	0.07	0.11	0.91	0.09	0.04
bottleneck	0.10	0.16	0.20	0.24	0.88	0.09

low similarity — high similarity

select strongest match per token

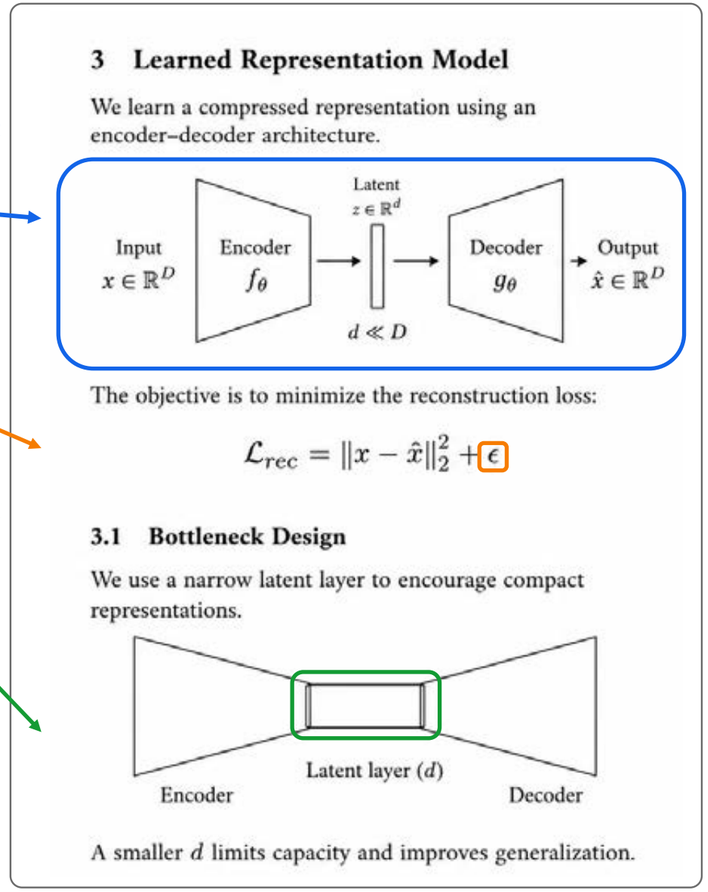
$$S(q,d) = \sum_i \max_j \langle q_i, p_j \rangle$$

semantic match

symbol match

visual/ concept match

Semantically matched page regions



Late interaction matches query tokens to relevant page regions.

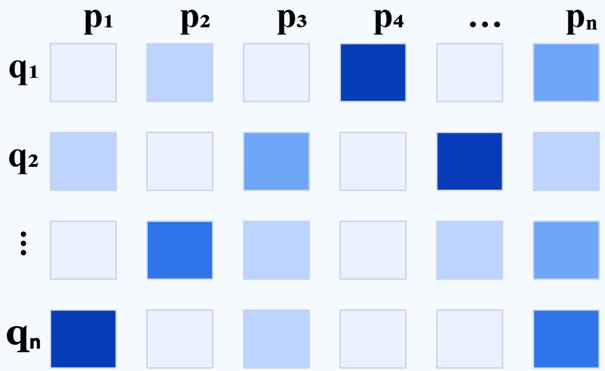
Challenge: Fine-Grained Matching, but Coarse Supervision

The retriever scores locally, but learns only from a page-level label.

INFERENCE: Local Interactions

Query tokens × Page patches

Many fine-grained interactions



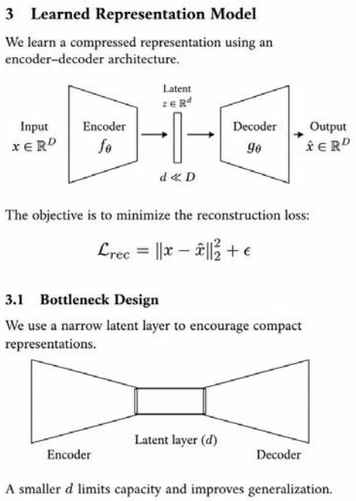
Nq × Nd local similarity scores

low similarity — high similarity

TRAINING: Page-Level Supervision

One page

One global relevance label



3 Learned Representation Model

We learn a compressed representation using an encoder–decoder architecture.

Input $x \in \mathbb{R}^D$ → Encoder f_θ → Latent $z \in \mathbb{R}^d$, $d \ll D$ → Decoder g_θ → Output $\hat{x} \in \mathbb{R}^D$

The objective is to minimize the reconstruction loss:

$$\mathcal{L}_{rec} = \|x - \hat{x}\|_2^2 + \epsilon$$

3.1 Bottleneck Design

We use a narrow latent layer to encourage compact representations.

Encoder — Latent layer (d) — Decoder

A smaller d limits capacity and improves generalization.

→ $y \in \{0, 1\}$

0 = not relevant

1 = relevant

WHAT THIS MEANS

The label tells whether the page is relevant, but not **where or why.**

The retriever may learn **shortcuts** rather than true **semantic alignment.**

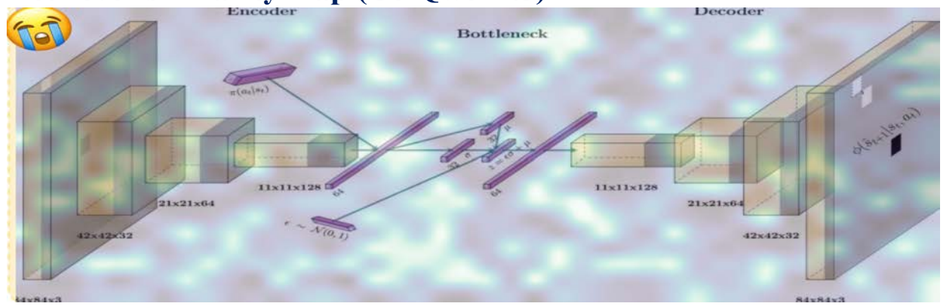
The missing piece is fine-grained supervision.

Motivation: Attention as a Scalable Local Signal

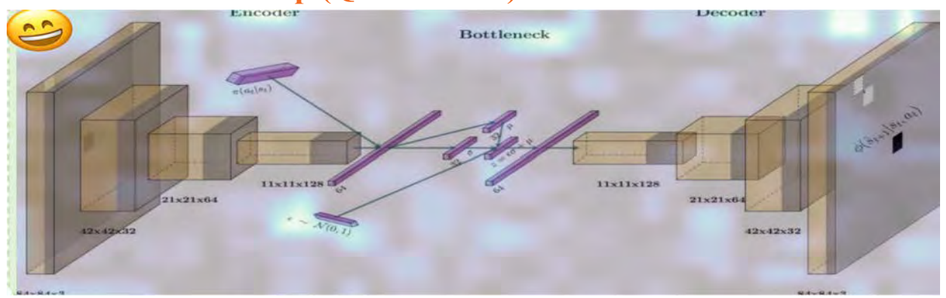
An MLLM can highlight query-relevant regions beyond exact keyword matches.

Query: What does the symbol ‘ε’ represent in the bottleneck section of the depicted autoencoder architecture?

Retriever similarity map (ColQwen2.5)



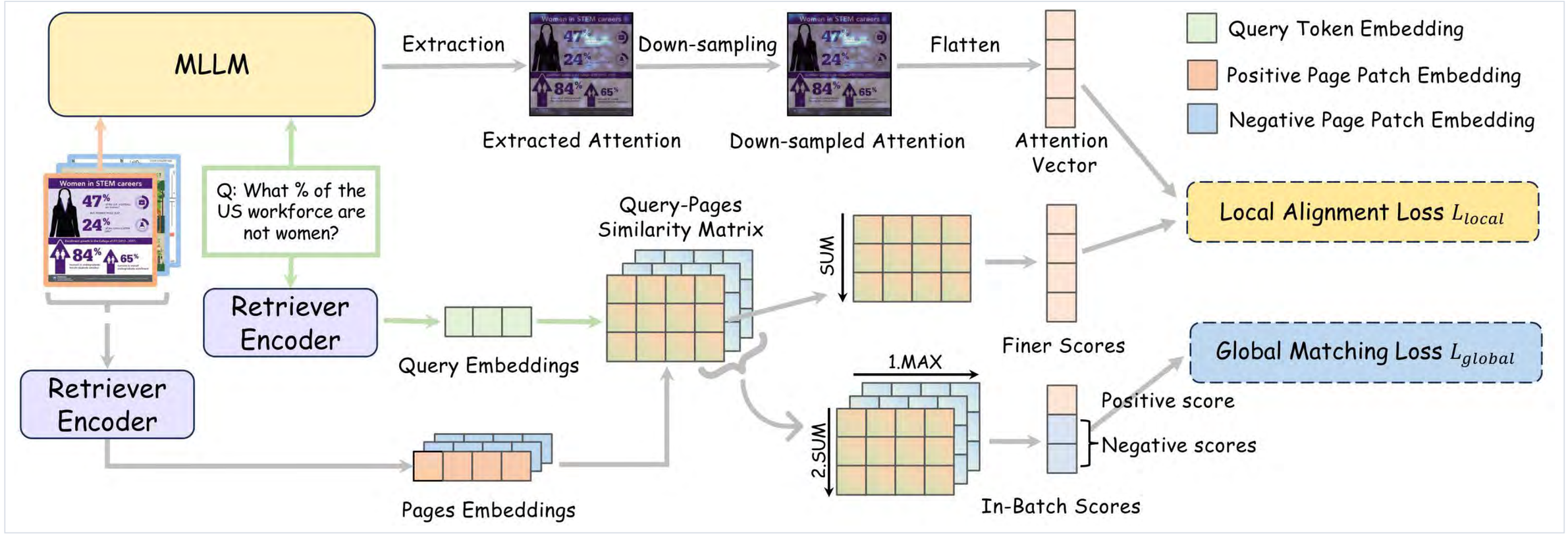
MLLM attention map (Qwen2.5-VL)



Evidence aspect	Retriever similarity map	MLLM attention map
Exact term: “bottleneck”	Captured	Captured
Small symbol: “ε”	Missed	Captured
Implicit relation: “autoencoder → encoder–decoder”	Partial	Captured

MLLM attention can serve as a scalable proxy for fine-grained supervision.

Method: AGREE in one view



1 **Query-Conditioned Attention Annotation** → 2 **Spatially Preserved Patch Alignment** → 3 **Global and Local Retriever Training**

> ***Stages 1& 2 construct the fine-grained supervision; Stage 3 trains the retriever.***

Stage 1: Query-Conditioned Attention Annotation

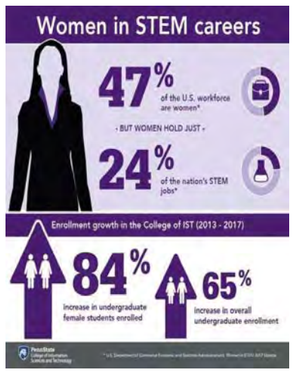
Use the frozen MLLM to extract where it attends for a query on a relevant page.

1 Query–Page Input

Query (natural language)

Q: What % of the US workforce are not women?

Page (screenshot)



2 All Query Tokens

Extract attention from **all query tokens** to image patches.

What | % | of | ••• | women | ?

Nq tokens

Why all tokens?

✓ Captures all pieces of evidence.

✓ Independent of answer generation.

3 Layer-wise Attention (within each layer)

For layer ℓ, each query token attends to **image patches**.

q_1 q_2 ⋮ q_{N_q}

Average over all query tokens within this layer

$A^{(\ell)}$

$$A^{(\ell)} = \frac{1}{N_q}\sum_i A^{(\ell,i)}$$

4 Multi-layer Aggregation

Layer 1 | Layer 2 | ••• | Layer L

Average over L layers (element-wise)

Final Attention Map

A

$$A = \frac{1}{L}\sum_i A^{(\ell)}$$

What it does

Extracts query-token attention over page patches and aggregates it across layers.

Why it matters

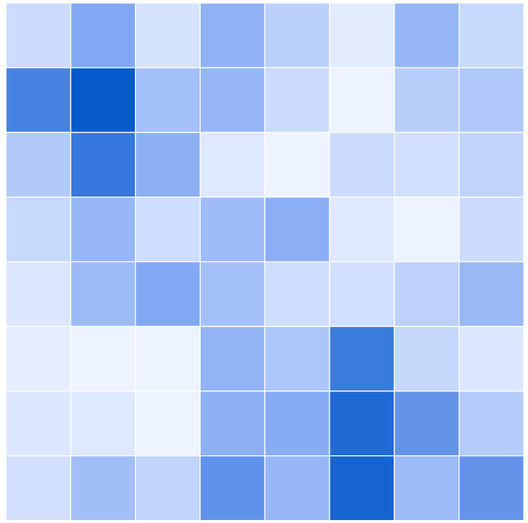
Makes the fine-grained supervision source explicit and transparent.

Stage 2: Patch-Level Attention Alignment

Align the MLLM attention to the retriever's patch grid.

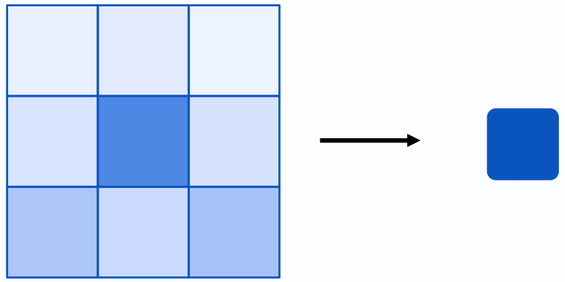
High-resolution Attention

(2048 patches)



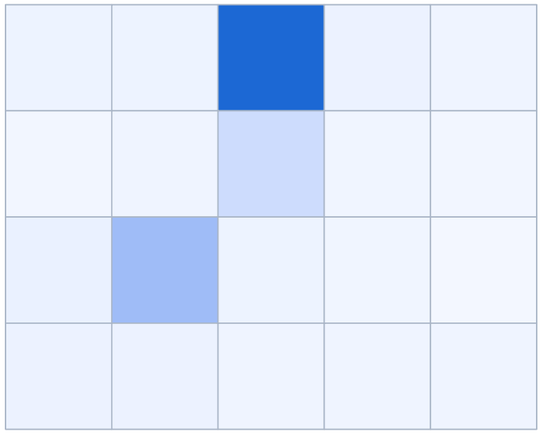
Adaptive Max Pooling

(keep the peak in each window)



Retriever Patch Grid

(1024 patches)



What it does

Aligns high-resolution MLLM attention with the retriever's patch grid.

Why it matters

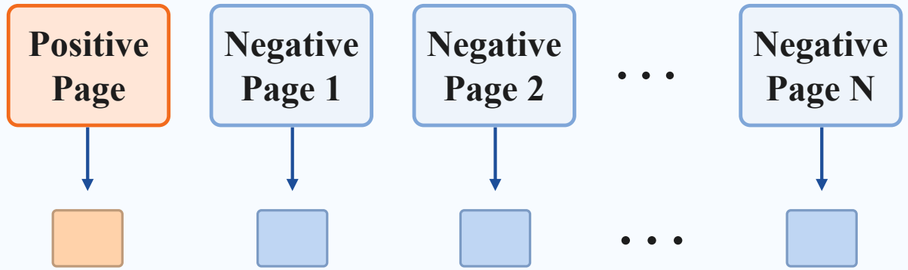
Preserves salient evidence (peaks) instead of averaging them away.

Stage 3: Global and Local Retriever Training

Train the retriever with both global relevance and local rationale supervision.

Global Ranking (In-batch)

Learn whether the page matches.

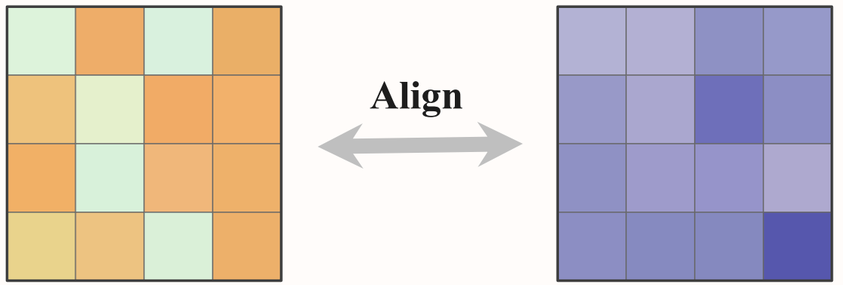


$$\mathcal{L}_{global} = \log\left(1 + \exp\left(\max_i S(q, d_i^-)\right) - \exp(S(q, d^+))\right)$$

Local Alignment (Positives only)

Learn which regions support relevance.

Retriever Patch Scores S **Target Attention A**



$$\mathcal{L}_{local} = 1 - \frac{S \cdot A}{\|S\|_2 \|A\|_2}$$

Total Loss: $\mathcal{L}_{total} = \mathcal{L}_{global} + \lambda \mathcal{L}_{local}$

What it does

Combines global ranking loss with local attention alignment.

Why it matters

Teaches the retriever both whether a page is relevant and where the evidence lies.

Experimental Setup

We evaluate **AGREE** across two benchmarks and two late-interaction retrievers.

Benchmarks

ViDoRe V2 **Implicit / non-extractive queries**

More challenging reasoning-oriented queries.

#Subsets	Domains	#Queries	#Pages	Pos./Query
7	ESG, Econ., BioMed.	10,201	1,240	2.4

- Generated non-extractive queries
- Often requires multi-region reasoning

ViDoRe V1 **More explicit matching**

Queries rely more on direct lexical overlap.

#Subsets	Domains	#Queries	#Pages	Pos./Query
6	Finance, Gov., Sci., Patent	10,000	1,200	1.6

- More extractive / explicit queries
- Easier lexical matching

Retriever Backbones

We implement **AGREE** on two strong late-interaction retrievers.

ColQwen2.5

Qwen2.5-VL backbone with strong reasoning ability.

ColPali (PaliGemma)

PaliGemma backbone with effective visual representation.

Retrieval Configuration

- Late interaction with MaxSim
- ColQwen2.5: max 768 patches / page
- ColPali: 1024 patches / page

Metrics

We follow the standard ViDoRe evaluation protocol.

nDCG@1

Whether the top-1 page is relevant

nDCG@5

Ranking quality within the top-5 results

Why nDCG@1 and nDCG@5?

- ✓ Standard metrics in ViDoRe and visual document retrieval
- ✓ Early-rank quality matters for downstream RAG

Main Results on ViDoRe V2

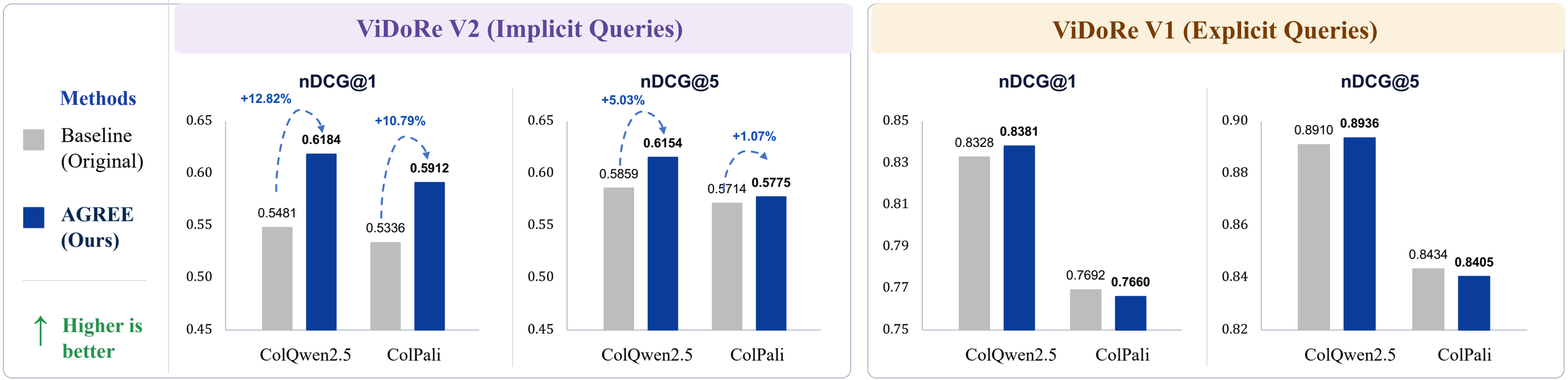
Detailed results adapted from the paper's original ViDoRe V2 table.

Model	ESG nDCG@1	ESG nDCG@5	Biomedical nDCG@1	Biomedical nDCG@5	Economics nDCG@1	Economics nDCG@5	ESG Human nDCG@1	ESG Human nDCG@5	Avg nDCG@1	Avg nDCG@5
Vision-Language Models										
BiCLIP	10.53	11.38	20.16	22.39	26.29	22.83	17.31	22.74	18.57	19.84
SigLIP (0-shot)	26.75	28.19	20.16	24.30	18.97	16.34	19.23	23.07	21.28	22.98
BiSigLIP	29.83	34.92	25.78	29.05	30.60	28.23	34.62	45.83	30.21	34.51
Multi-modal Large Language Models										
DSE	47.37	55.15	52.66	55.29	62.50	52.80	57.69	61.74	55.06	56.25
ColPali	51.75	54.05	52.03	57.56	55.17	53.30	54.49	63.67	53.36	57.14
ColQwen2.5	51.75	57.12	56.56	60.64	52.59	52.90	58.33	63.70	54.81	58.59
Ours										
AGREE$_{pali}$	**57.46†**	**58.87†**	**52.03**	**55.51**	**59.05†**	**51.33**	**67.95†**	**65.28†**	**59.12**	**57.75**
AGREEQwen2.5	**60.09***	**58.43**	**56.41**	**60.49**	**59.05***	**56.50***	**71.80***	**70.73***	**61.84**	**61.54**

† : Significant improvement over ColPali; * : Significant improvement over ColQwen2.5.

Takeaway | **AGREE** achieves the **best average results** on ViDoRe V2; gains are especially strong with **Qwen2.5** and remain consistent across **multiple subsets.**

Generalization Across Base Retrievers and Benchmarks



Takeaway

AGREE shows strong gains on ViDoRe V2; on ViDoRe V1 it remains competitive, with small Qwen gains and slight Pali drops.

How Should We Align Retriever Scores with MLLM Attention?

Local Alignment Objectives

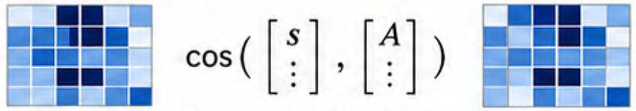
1 KL-100%: Full-distribution matching

Match all patches after softmax, including low-attention regions.

2 KL-10% / KL-20%: Salient-region matching

Supervise only the top 10% / 20% attended patches.

3 Cosine-100%: Pattern alignment



Align score pattern with attention pattern, no threshold needed.

Controlled Results on ViDoRe V2 (ColQwen2.5)

Global-only baseline: nDCG@1 = 0.5481, nDCG@5 = 0.5859

Local Objective	What it aligns	nDCG@1	nDCG@5
KL-100%	Full distribution	0.5902	0.5993
KL-10%	Top 10% salient patches	0.6037	0.6078
KL-20%	Top 20% salient patches	0.6086	0.6103
Cosine-100%	**Directional pattern**	**0.6184**	**0.6154** ★ Best

Why Cosine-100% Works Best

- **Threshold-free**
- **Less sensitive to noisy background**
- **Focuses on salient pattern agreement**

$$L_{\cos} = 1 - \frac{\langle s, A \rangle}{\|s\| \ \|A\|}$$

same local loss used in Method

Takeaway

Fine-grained attention supervision consistently improves over global-only training; cosine gives the most robust local alignment.

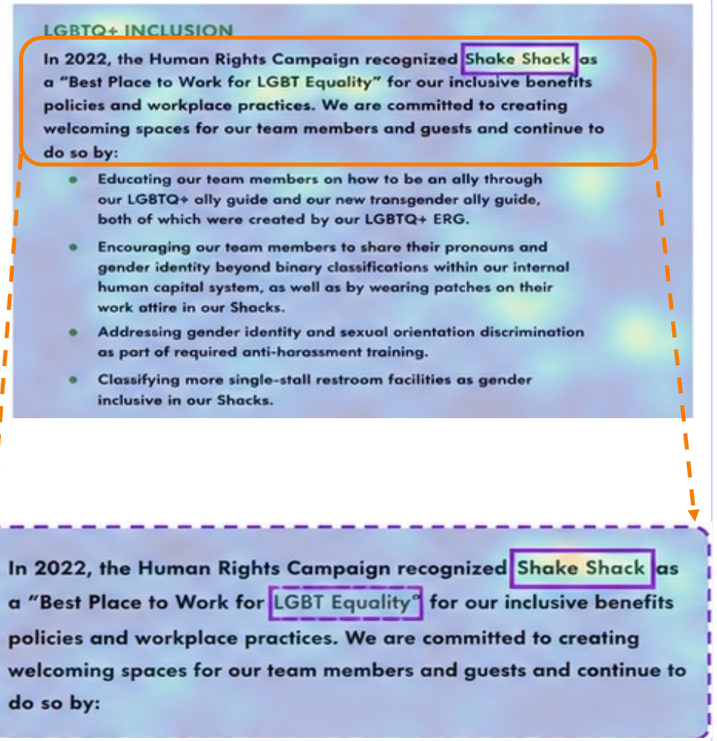
Case Study: AGREE Captures Implicit Semantic Evidence

Query: What steps has **Shake Shack** taken to create a welcoming environment for **gay** customers?

Original document page

Focus region

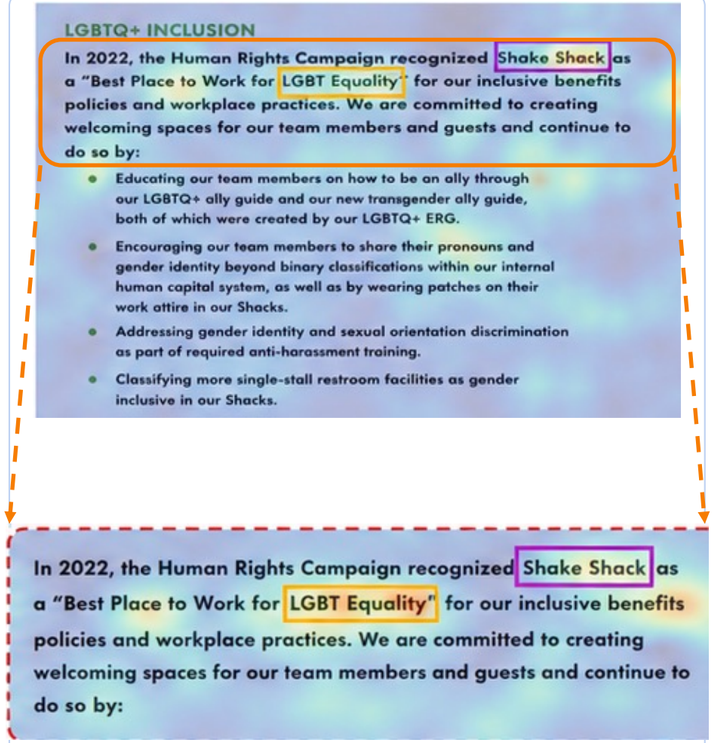
Baseline (ColQwen2.5) Similarity Map



✗ Focuses mainly on the entity **"Shake Shack".**

VS.

AGREE (Ours) Similarity Map



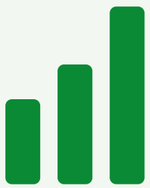
✓ Also captures **"LGBT Equality"** as implicit semantic evidence.

Takeaway

AGREE captures the **implicit correspondence** between **"gay"** and **"LGBT Equality"**, beyond the entity match on **"Shake Shack".**

AGREE:

1 **Query-conditioned MLLM attention** provides fine-grained supervision for visual document retrieval, ***without manual region annotations***



2 AGREE achieves a ***+12.82% relative nDCG@1 improvement*** on ViDoRe V2, while retaining standard MaxSim inference and requiring ***no MLLM at deployment.***

Code: https://www.github.com/VickiCui/AGREE

Thank you & QA