Title Slide

- Title: MINDER: Multi-View Document Representation for Generative Retrieval
- Authors: [Yongqi Li , Nan Yang , Liang Wang , Furu Wei , Wenjie Li]
- Presented by: [SURBHI SHARMA]
- Date: [Insert Date]

Introduction

Challenge: How to retrieve passages without relying on dense encoders?

MINDER: A generative retrieval model using multi-view semantic identifiers.

Focus: Representing and retrieving documents through natural language strings.

Motivation

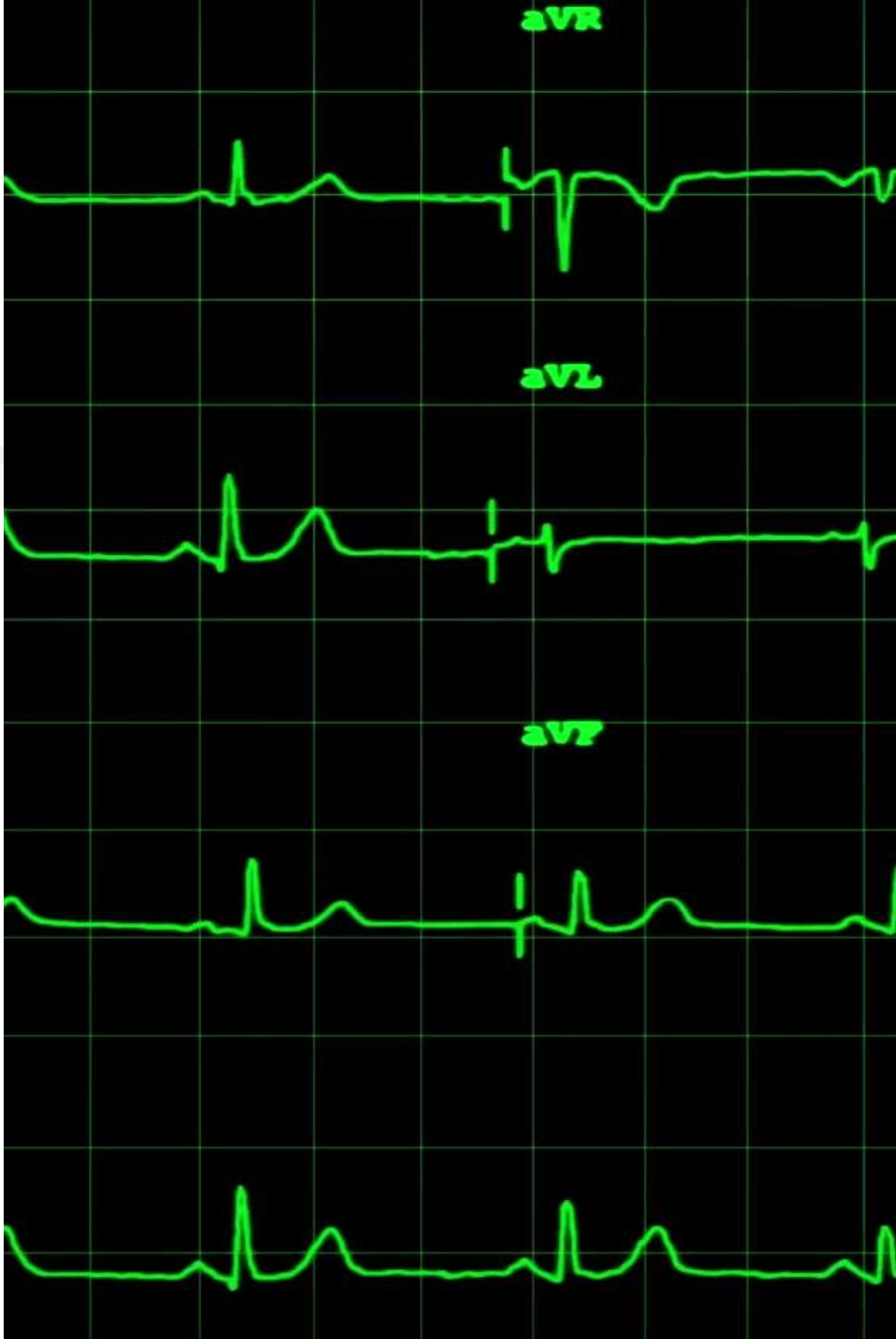
Numeric IDs (e.g., DSI) are hard to learn and lack meaning

Need semantic and diverse identifiers

MINDER introduces three views: titles, substrings, pseudo-queries

Identifier Views

- Title View (t): Wikipedia page titles
- Substring View (S): Meaningful text spans from the passage
- Pseudo-Query View (Q): Queries generated from the passage



Pseudo-Query Generation

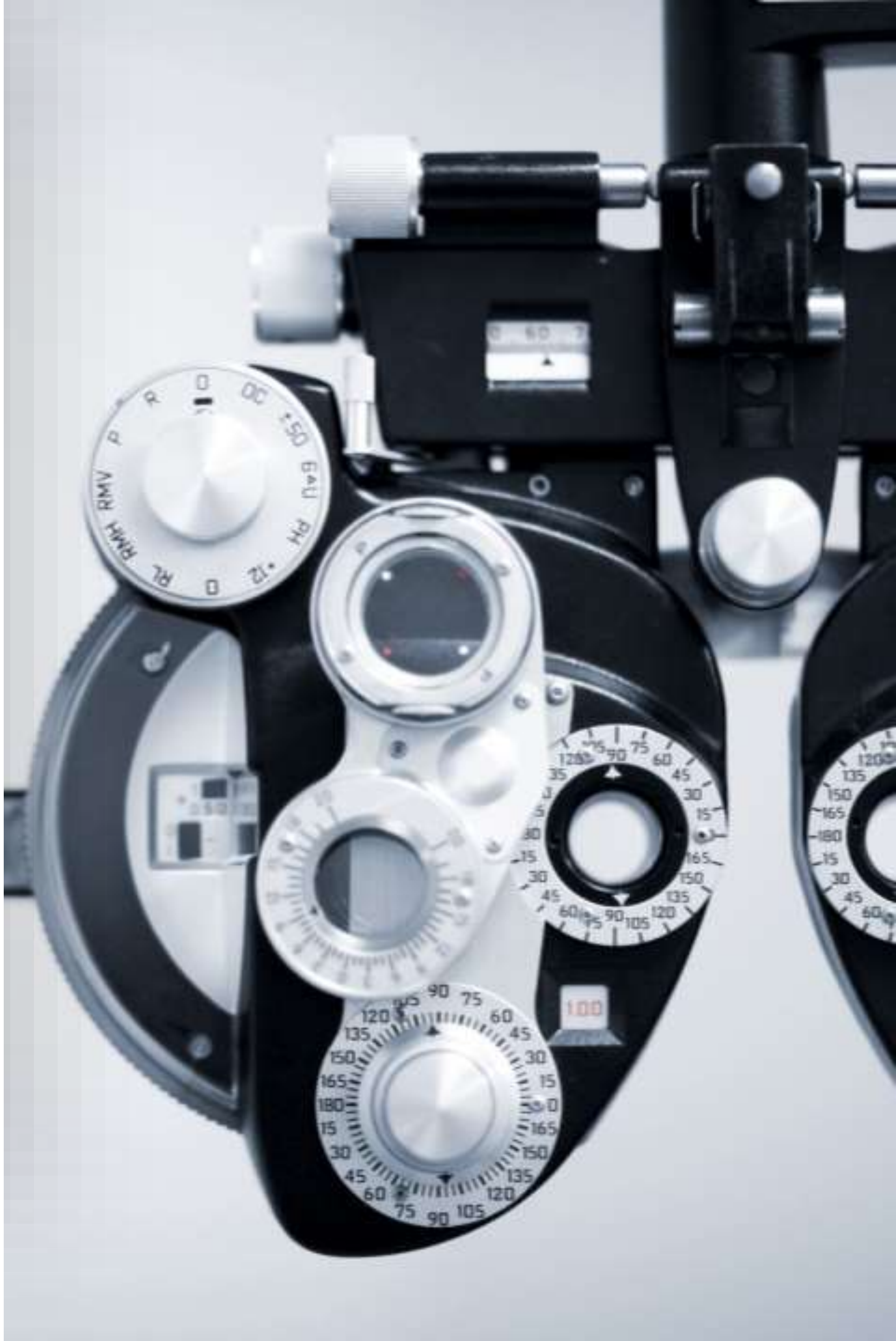
- Trained a Query Generation Model (QG) using labeled query-passage pairs
- Generated diverse queries using top-k sampling
- For each passage, produce a set of pseudo-queries

Model Training (AM)

- Used autoregressive models (BART/T5)
- Trained to generate identifiers conditioned on the input query
- Training samples:
- - Input = [prefix; query]
- - Target = identifier (title, substring, or pseudo-query)

Constrained Generation using FM-Index

- FM-index allows efficient search and valid decoding
- Flatten identifiers with special tokens (e.g., <TS>, <TE>, <QS>, <QE>)
- Beam search constrained to generate only valid identifiers



Retrieval and Ranking

For a query, generate candidate identifiers:

- Tg (titles), Sg (substrings), Qg (pseudo-queries)

Use FM-index to match identifiers to passages

Use language model scores for ranking

Datasets Used

- Natural Questions (NQ) and TriviaQA: Wikipedia-based
- MSMARCO: Web-based queries and passages
- Passages ≤100 words, multiple per document

Baselines

- Generative: DSI, SEAL
- Dual-Encoder: DPR, GAR
- Lexical: BM25
- MINDER outperforms most on NQ and ties on TriviaQA

Comparison of Retrieval Methods and MINDER

Method	Type	Indexing	Key Strength	Weakness
BM25	Lexical (Sparse)	Inverted Index	Fast, interpretable	Ignores semantics
DPR	Dense Retrieval	Dual encoder, vector index	Semantic matching	Hard to represent multi-view queries
GAR	Dense (Adversarial)	Dual encoder + Generator	Learns hard negatives dynamically	More complex training
SEAL	Generative	Identifiers + FM-index	Combines generative retrieval with indexing	Limited identifier types
DSI-BART	Generative	Encoded in LM weights (no corpus)	No external index needed	Doesn’t scale well, static index
MINDER	Generative + Index	Multiview identifiers + FM-index	Best of all worlds: titles, substrings, queries	Manual scoring heuristic

What Are SEAL-LM and SEAL-LM+FM?

SEAL-LM:
Uses a language model to generate identifiers but doesn't store them in a retrievable index. Retrieval happens purely through generation, not search.

✓ Semantic

✗ Slow, doesn't scale

SEAL-LM + FM:
Adds an FM-index over the generated identifiers (like BWT structure) for faster lookup.

✓ Efficient lookup

✗ Still limited identifier diversity

Dataset	Metric	DPR (Dense)	MINDER (Generative)
NQ	@5	**68.3**	65.8
	@20	**80.1**	78.3
	@100	86.1	**86.7**
TriviaQA	@5	72.7	**68.4**
	@20	**80.2**	78.1
	@100	**84.8**	**84.8**

? Interpretation:

On **Natural Questions (NQ)**:

- **DPR is slightly better** on hits@5 and @20.
- **MINDER wins** at hits@100, indicating **better long-tail recall**.

On **TriviaQA**:

- **DPR has the edge on hits@5 and @20**, but
- **Both tie at hits@100**.

Implementation Details

Backbone: BART-large

Data ratio (Title:Substring:PQ) = 3:10:5

Optimizer: Adam, LR=3e-5, 800k updates

Framework: fairseq, 8×V100 GPUs

Inference Example

Query: "Who sings 'Does He Love You'?"

AM generates: pseudo-query → matched using FM-index

Ranked passages retrieved via constrained beam search

Evaluation (Hits@k)

- Metric: Hits@5, 20, 100
- MINDER highest on NQ (Hits@100: 86.7%)
- Slightly lower than DPR on TriviaQA
- Shows effectiveness of multi-view semantic identifiers

Unsupervised Data

Added pseudo-query and span-based unsupervised examples

Improves generalization to noisy queries

MINDER: +1.2 points Hits@5; SEAL: +2.3 points

Method Variant	Hits@5 (NQ)
MINDER (no extra data)	64.6%
+ Span as Query	65.9%
+ Pseudo-query	65.8%

Why Not Numeric Identifiers?

Hard to learn

Lack semantic grounding

MINDER focuses on textual identifiers for interpretability and better training

Key Takeaways

Multi-view identifiers boost generative retrieval

FM-index enables valid and efficient decoding

Outperforms prior generative approaches

Useful for low-supervision and open-domain QA settings

Ablation Study – Role of Identifier Views

- To answer 3 key questions:
- Do all three identifier views (Title, Substring, Pseudo-Query) contribute meaningfully?
- How do they help across datasets like NQ and MSMARCO?
- Is there a difference when using numeric vs. semantic identifiers?

Removed View	Hits@5 Drop (NQ)
Pseudo-Query Removed	–2.4%
Substring Removed	–2.7%
Title Removed	–1.9%

- **Different identifier views dominate in different datasets:**
  - Substring > Title > Pseudo-query for **NQ**
  - Pseudo-query > Substring > Title for **MSMARCO**
- ? Interpretation:
- **Substrings** help more when query is factual and content-rich (e.g., NQ)
- **Pseudo-queries** help more when query phrasing is diverse (e.g., MSMARCO)

- **Beam Size Sensitivity:**
- Varying beam sizes (5–20) affects performance slightly, more so on TriviaQA than MSMARCO.
- MINDER is robust across beam settings.

Aspect	Dense Retrieval	Generative Retrieval (MINDER)
Query-Document Interaction	Via embeddings (dot product)	Via decoder likelihood
Strengths	Fast, scalable, robust	Semantic, flexible, interpretable
Weaknesses	Limited by vector representation	Error-prone generation, slower
Use case	Large-scale search	Few-shot, complex queries, explainability

Bottom Line:

- **MINDER is strong** in **semantic flexibility**, especially with limited supervision or in few-shot settings.
- **Dense retrieval is strong** in **speed and robustness**, especially for large-scale applications.
- **Combining both** (e.g., MINDER for candidate generation + dense reranker) could yield **best of both worlds**

Limitations & Future Work

- Slower inference vs dual-encoders
- Relies on identifier quality
- Future: hybrid retrievers, better identifier design

✓ **Conclusion**

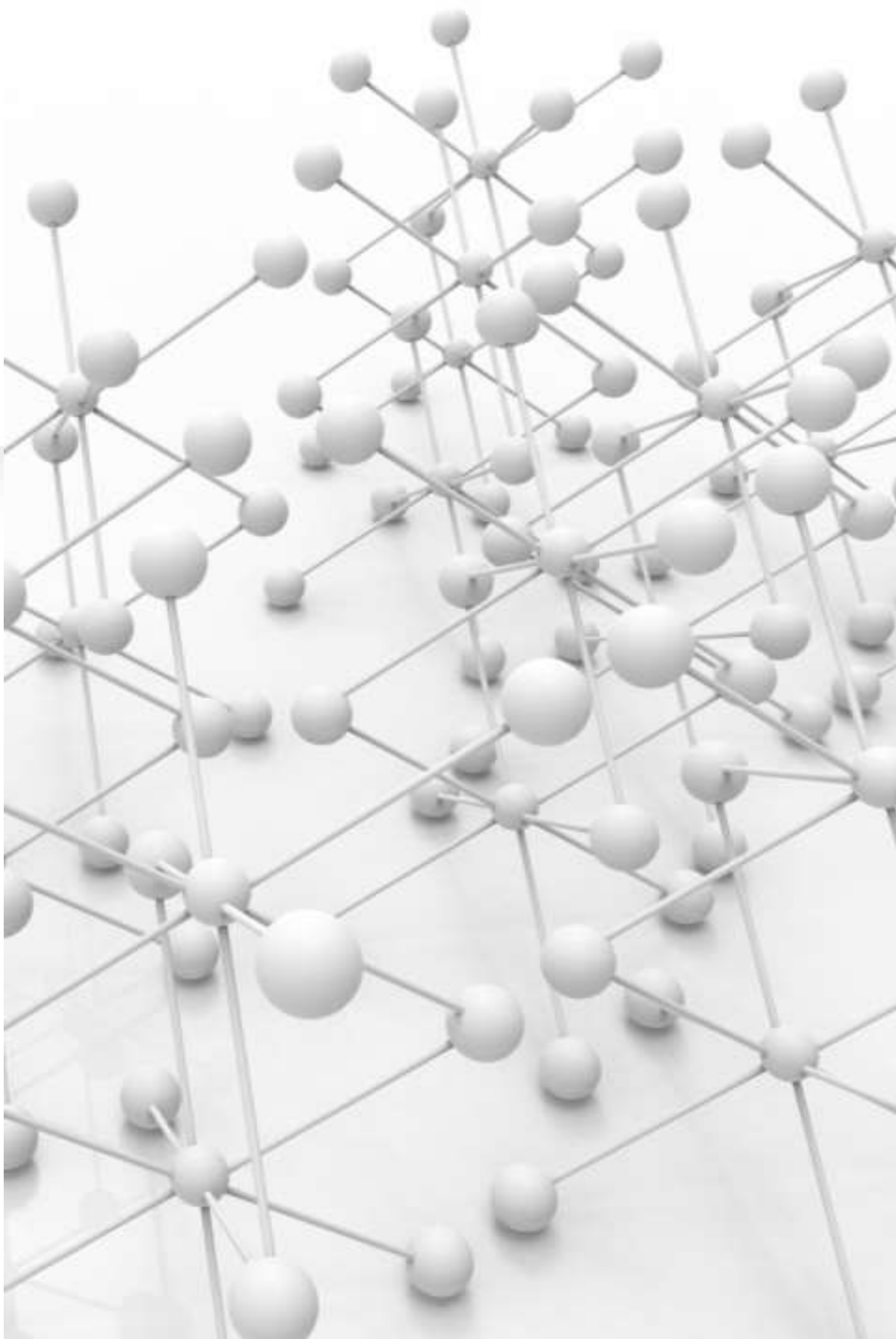
- **MINDER** combines an **autoregressive language model** with **multi-view identifiers**:
  - **Titles**, **substrings**, and **pseudo-queries**
- These identifier views **complement** each other, boosting retrieval performance.
- MINDER achieves **state-of-the-art** results among generative retrievers on datasets like **NQ**, **TriviaQA**, and **MS MARCO**.

Neural Ranker Integration:

- Replace heuristic identifier scoring with a **learned scoring model** for adaptive weighting.
- **Model Architecture**:
- Explore **bidirectional or non-causal models** (e.g., **T5** or **encoder-decoder** models) to overcome **BART's left-to-right limitation**.

New Settings:

- Apply to **few-shot** and **domain-specific** retrieval scenarios.
- **Memory Optimization**:
- Reduce identifier storage cost while maintaining fast access (e.g., enhanced indexing structures).



Thank You!

Questions?