中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

中国科学院大学
University of Chinese Academy of Sciences

MMDBench

A Benchmark for Hybrid Query in Multimodal Database

Along Mao[1,2] , Chuan Hu[1,2] , Chong Li[1] , Huajin Wang[1] , Junjin Rao[1,2], Kainan Wang[1,2], Zhihong Shen[1]

[1] Computer Network Information Center, Chinese Academy of Sciences

[2] University of Chinese Academy of Sciences

Along Mao | Bench 2023

Outline

Part 1. Background

Part 2. Problem

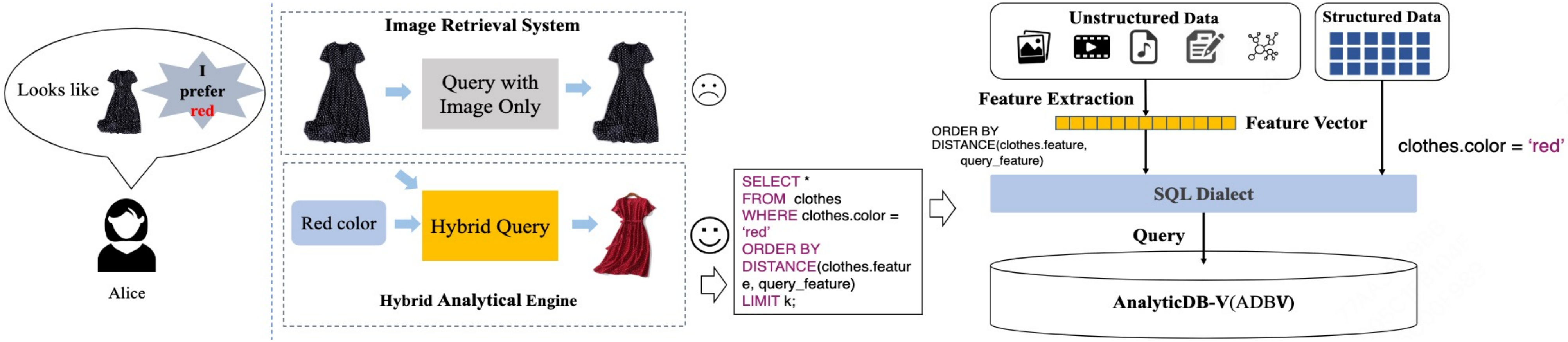
Part 3. Proposed Solution

Part 4. Experiment

Part 5. Conclusion

Background

- Instructed data occupies a huge proportion of Internet and scientific data[1-3]
- Data presents a variety of modalities, and semantic information needs to be mined[4-6]
- Some database systems try to provide solutions for multimodal data hybrid queries[7-8]



```
SELECT *
FROM clothes
WHERE clothes.color =
'red'
ORDER BY
DISTANCE(clothes.featur
e, query_feature)
LIMIT k;
```

Fig .1. Example for Hybrid Query[8]

Traditional methods (image recognition systems) cannot take into account structured query requirements, e.g., items with red color.

➡

Joint structured/unstructured query (AI+DB) filters for both structured features (color) and unstructured features

1. "Structured vs. Unstructured Data". www.datamation.com. Retrieved 2018-10-02.
2. "What is unstructured data?". https://www.mongodb.com/unstructured-data. Retrieved 2021-2-25.
3. John Gantz and David Reinsel. 2011. Extracting value from chaos. IDC iview 1142, 2011 (2011), 1–12.
4. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In The semantic web. Springer, 722–735.
5. Sebastián Ferrada, Benjamin Bustos, and Aidan Hogan. 2017. IMGpedia: a linked dataset with content-based analysis of Wikimedia images. In International Semantic Web Conference. Springer, 84–93
6. Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM 57, 10 (2014), 78–85.
7. ZHAO Z, SHEN Z, MAO A, et al. PandaDB: An AI-Native Graph Database for Unified Managing Structured and Unstructured Data[J].
8. WEI C, WU B, WANG S, et al. AnalyticDB-V: a hybrid analytical engine towards query fusion for structured and unstructured data[J/OL].

Problem

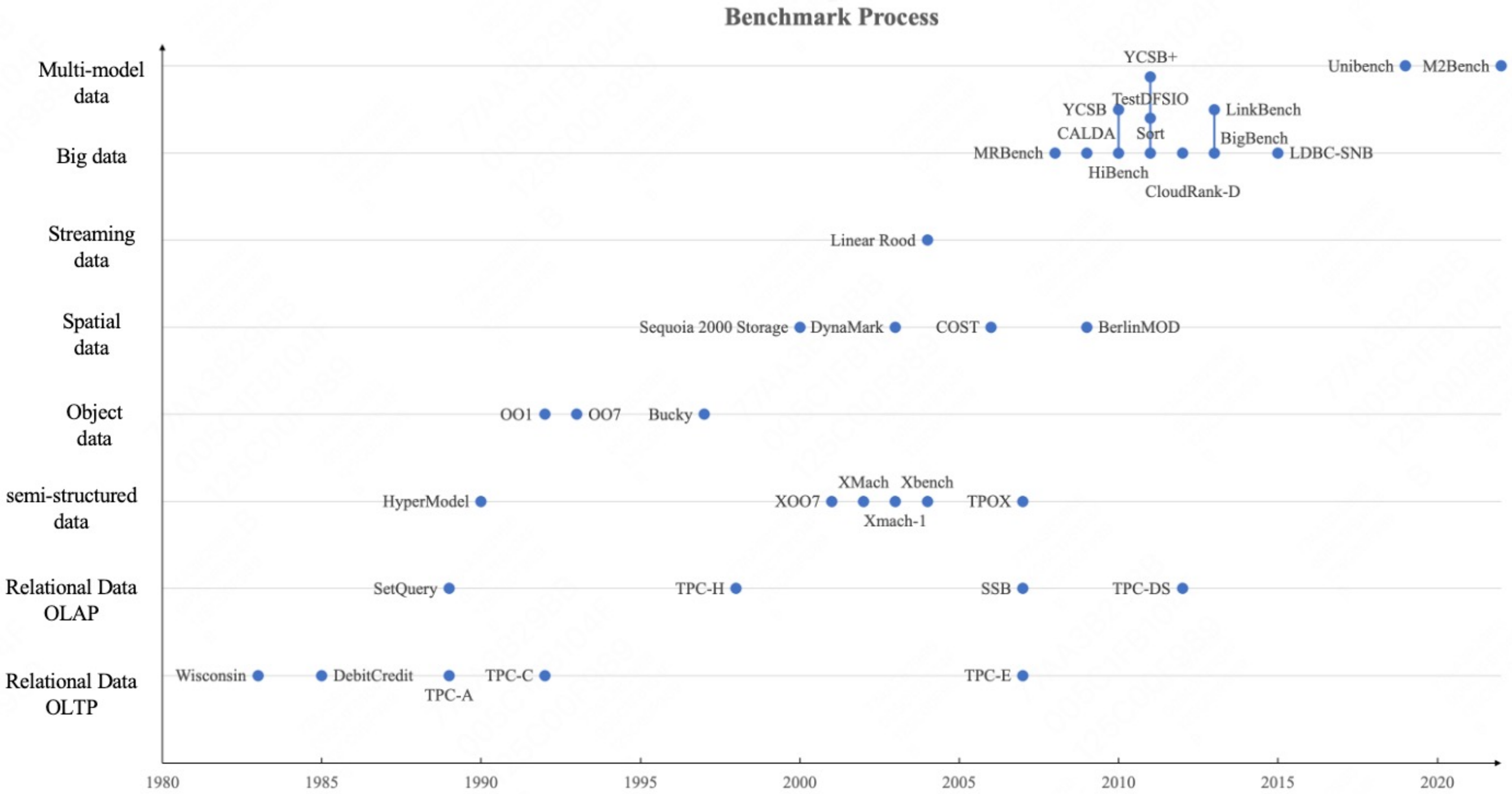


Fig.2. Benchmark Process

The importance of benchmarks:

a. Performance Evaluation

b. Innovation and Progress Tracking

c. Quality Assurance

d. System Optimization

e. Decision-Making

There is a gap in the evaluation benchmark for **multimodal data**, and the related research community is in urgent need of a set of fair and objective evaluation benchmarks to simplify the process of comparing different database

Proposed Solution

- **Multimodal Data Simulation Methodology**
  - Based on tools and datasets
  - Leveraging real-world distribution patterns
  - Controllable scale
- **Hybrid Query Workload Design**
  - Aligning with real-world application scenarios
  - Featuring typicality, interpretability, and Portability
  - Inspired by the key operation and choke point
  - Incorporating both structured and unstructured data for collaborative retrieval
- **Universal Benchmark Framework**
  - Facilitating quick integration of benchmark
  - Developing a plugin-based architecture
  - Query and storage operations are abstracted into CRUD

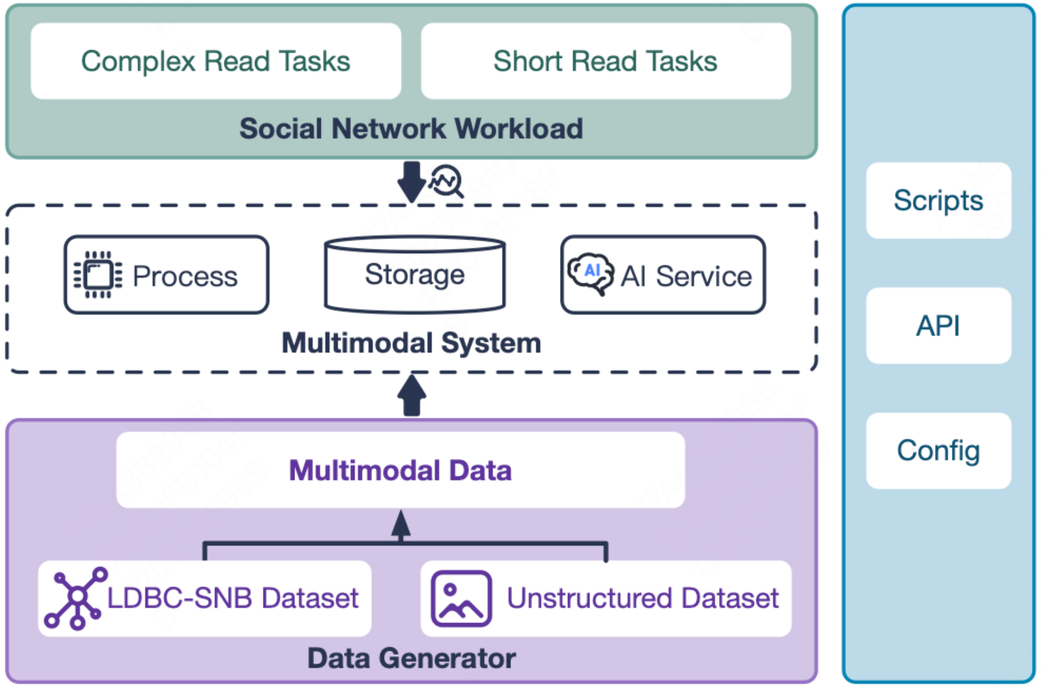
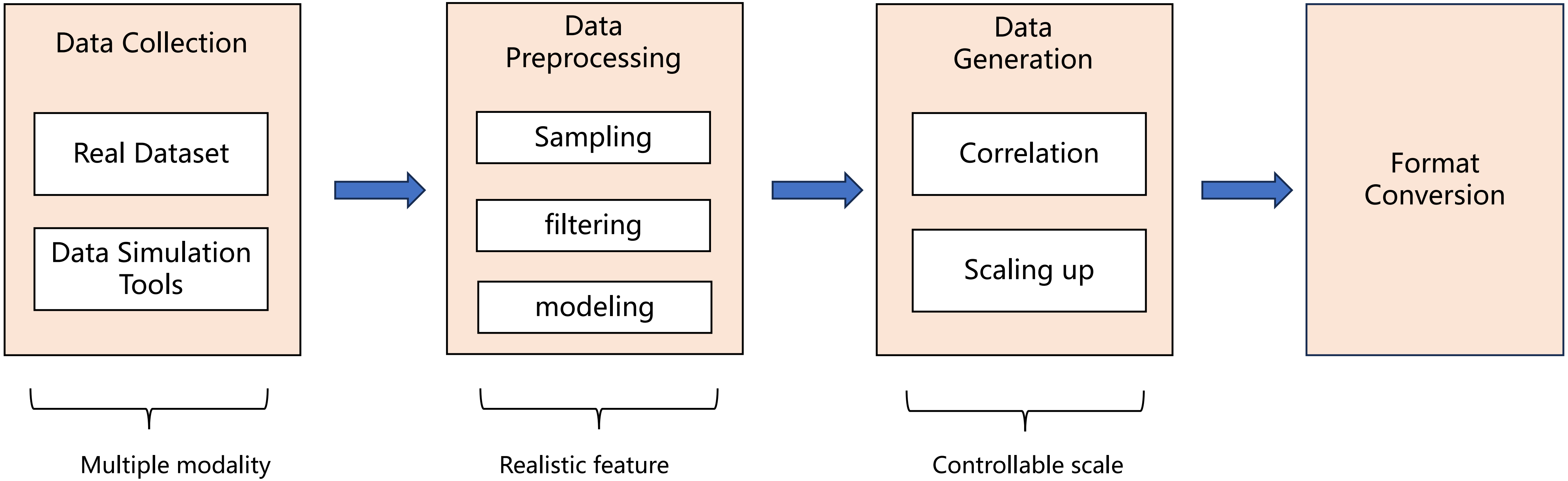


Fig.3. Overview Of MMDBench

Data Simulation Methodology



Data Collection

Real Dataset

Data Simulation Tools

Data Preprocessing

Sampling

filtering

modeling

Data Generation

Correlation

Scaling up

Format Conversion

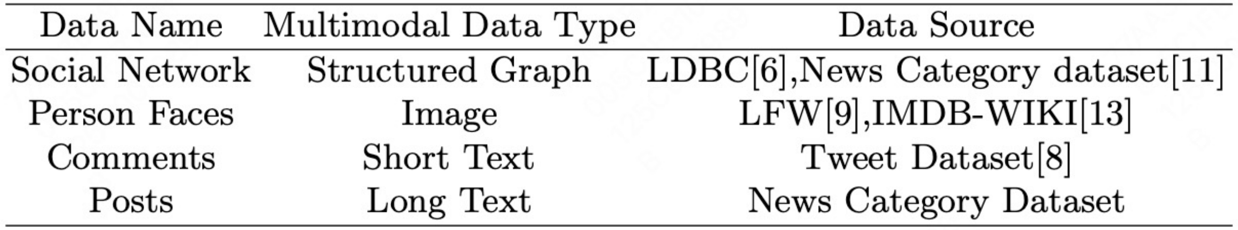
Multiple modality

Realistic feature

Controllable scale

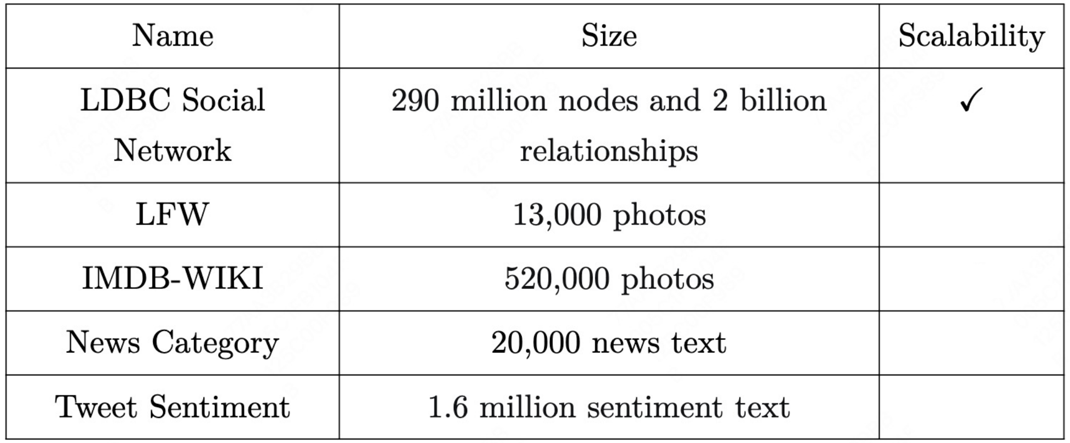
Multimodal Data Collection

Table 1. Dataset in MMDBench



Data Name	Multimodal Data Type	Data Source
Social Network	Structured Graph	LDBC[6],News Category dataset[11]
Person Faces	Image	LFW[9],IMDB-WIKI[13]
Comments	Short Text	Tweet Dataset[8]
Posts	Long Text	News Category Dataset

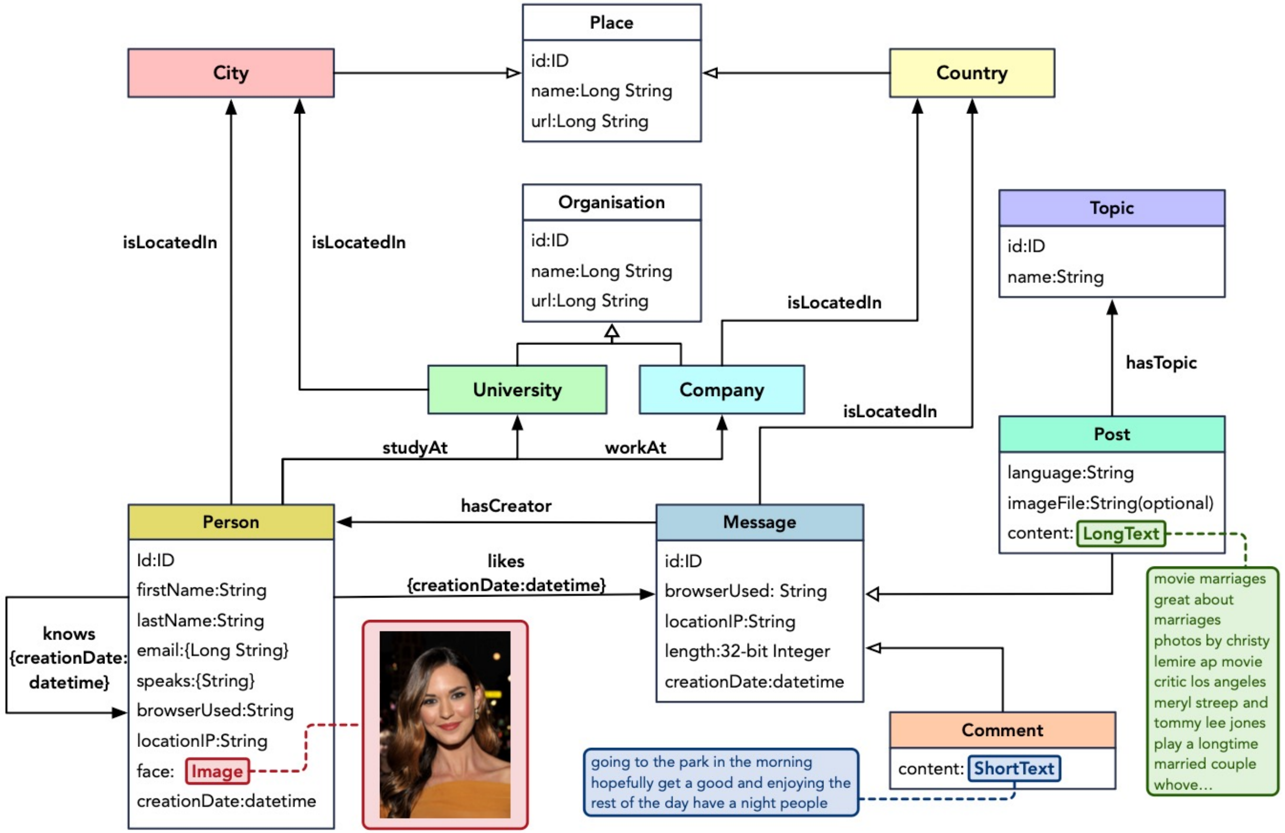
Table 2. Dataset Characteristics



Name	Size	Scalability
LDBC Social Network	290 million nodes and 2 billion relationships	✓
LFW	13,000 photos	
IMDB-WIKI	520,000 photos	
News Category	20,000 news text	
Tweet Sentiment	1.6 million sentiment text	

Modeling & Correlation

Modeling based on extend property graph： Node, Relationship, Structured Property and Unstructured Property.



Place
id:ID
name:Long String
url:Long String

City

Country

Organisation
id:ID
name:Long String
url:Long String

Topic
id:ID
name:String

isLocatedIn

isLocatedIn

isLocatedIn

isLocatedIn

hasTopic

University

Company

studyAt

workAt

Post
language:String
imageFile:String(optional)
content: LongText

Person
Id:ID
firstName:String
lastName:String
email:{Long String}
speaks:{String}
browserUsed:String
locationIP:String
face: Image
creationDate:datetime

hasCreator

likes
{creationDate:datetime}

Message
id:ID
browserUsed: String
locationIP:String
length:32-bit Integer
creationDate:datetime

knows
{creationDate:
datetime}

Comment
content: ShortText

going to the park in the morning hopefully get a good and enjoying the rest of the day have a night people

movie marriages great about marriages photos by christy lemire ap movie critic los angeles meryl streep and tommy lee jones play a longtime married couple whove…

Fig.4. Multimodal Social Network Schema

Data correlation

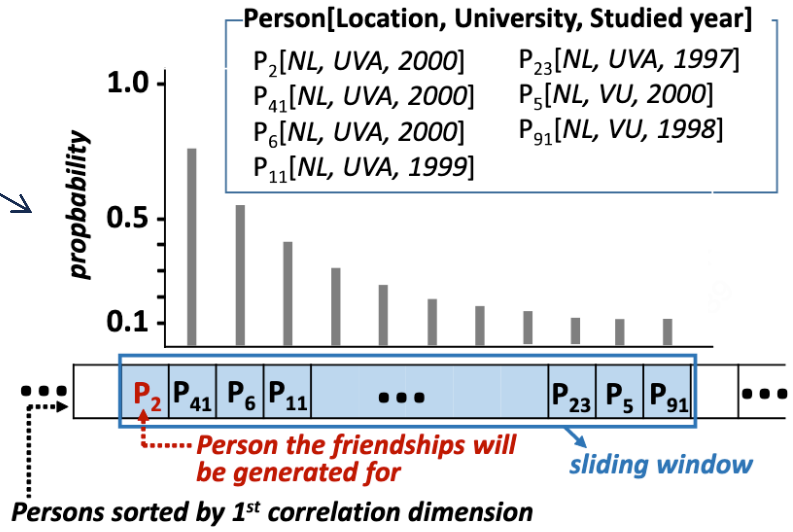
Structured Data → Structured Property

Unstructured Data → Unstructured Property

Storage: storing raw data information for Unstructured data

Scaling Data

- Text Data[9]:
  - P% Probability Random Swap
  - Q% Probability Synonym Replacement
  - M% probability Random Insertion
  - N% probability Random Deletion
- Image Data:
  - Strategy 1: Random Sampling Method Based on Large-Scale Data
  - Strategy 2: Scaling Data Based on Image Generative Algorithms
- Graph Data[10]:
  - Person Generation (Based on Dictionary)
  - Generation of Relationships
    - Sliding Window Algorithm Based on Multidimensional Correlation Ranking
    - Node Degree Selection Algorithm Based on Power Law Distribution



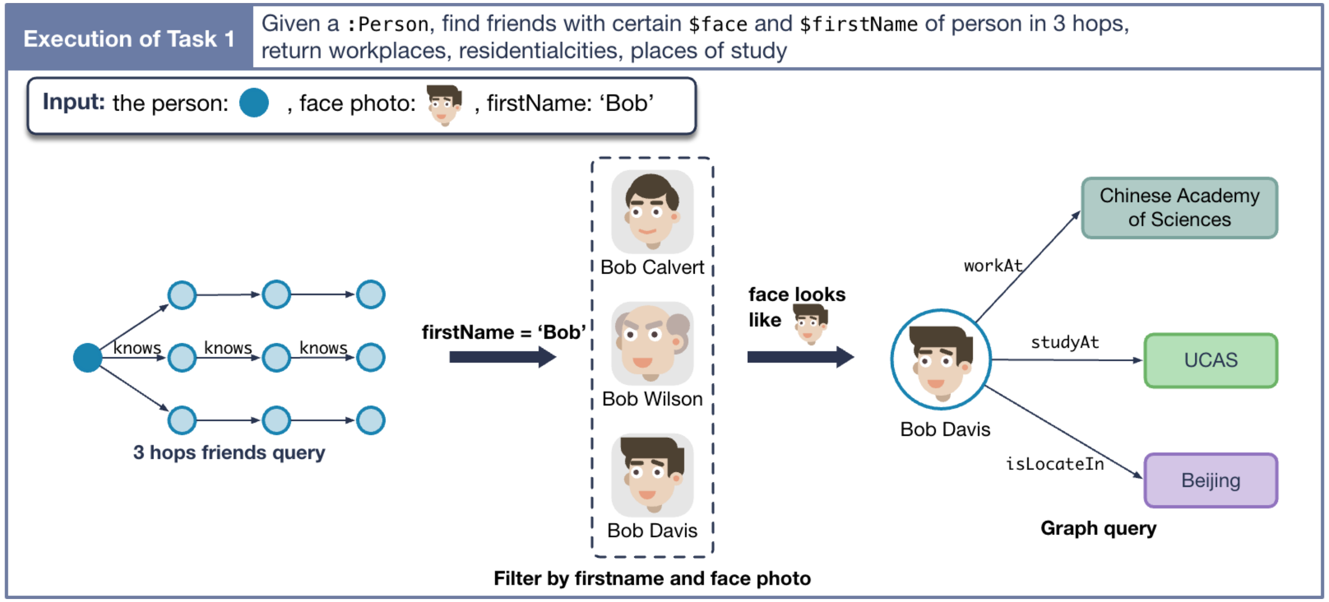
[9] Wei J, Zou K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks[J]. arXiv preprint arXiv:1901.11196, 2019.
[10]Erling O, Averbuch A, Larriba-Pey J, et al. The LDBC social network benchmark: Interactive workload[C]//Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. 2015: 619-630.

Workload

- **Scenario**： Social Network
- Choke-point based design
- Collaborative Retrieval of Structured and Unstructured Data
- Each task involves at least two modalities of data

Table 3. Key operations of MMDBench

Data type	Operation
Structured Graph Data	Join Selection Aggregation Pattern Matching Shortest Path
Unstructured Data	Unstructured Property Filtering Relationship Inference Similarity Matching



Execution of Task 1

Given a :Person, find friends with certain $face and $firstName of person in 3 hops, return workplaces, residentialcities, places of study

Input: the person: ● , face photo: , firstName: 'Bob'

Fig.5. Example of Workload

Table 4. Tasks of MMDBench

	Task	Operation
complex read	T1	Structured and unstructured property filtering
	T2	Multiple unstructured property filtering
	T3	Hybrid query with join
	T4	Hybrid query with aggregation
	T5	Hybrid query with Subgraph Matching
	T6	Relationship inference
	T7	Hybrid query with unweighted shortest path
short read	T8	Face recognition and pattern matching
	T9	Face recognition and pattern matching
	T10	Sentiment analysis
	T11	Sentiment analysis and pattern matching

Benchmark Framework

- **Structured Data**:
  - Storage：Graph Databases, Relational Databases, etc.
  - Key Elements: Node, Relationship, Property
  - Abstract CRUD Interfaces for Node/Relationship Storage and Query Operations

- **Unstructured Data**:
  - Storage: Object Storage System(OSS)/File System
  - Query：External AI services with plug-in architecture
  - Ability to analyze data from different modalities:
    - Image: Face Recognition
    - Short Text: Sentiment Analysis
    - Long Text: News Classification/Topic Extraction

- **Coordination Client**:
  - Facilitating communication and interaction among various systems in multi-modal queries.

Experiment

Table 5. Characteristics of dataset

SF	Number					Import Time(ms)	Generator Time(ms)
	Person	Post	Comment	Likes	Has_Topic		
1	10,295	1,121,226	1,739,438	1,870,268	672,735	18,329	197,052
3	25,066	2,873,419	5,343,582	6,244,522	1,724,051	37,155	264,788
5	31,505	3,665,392	7,041,356	8,468,619	2,199,235	39,920	331,963

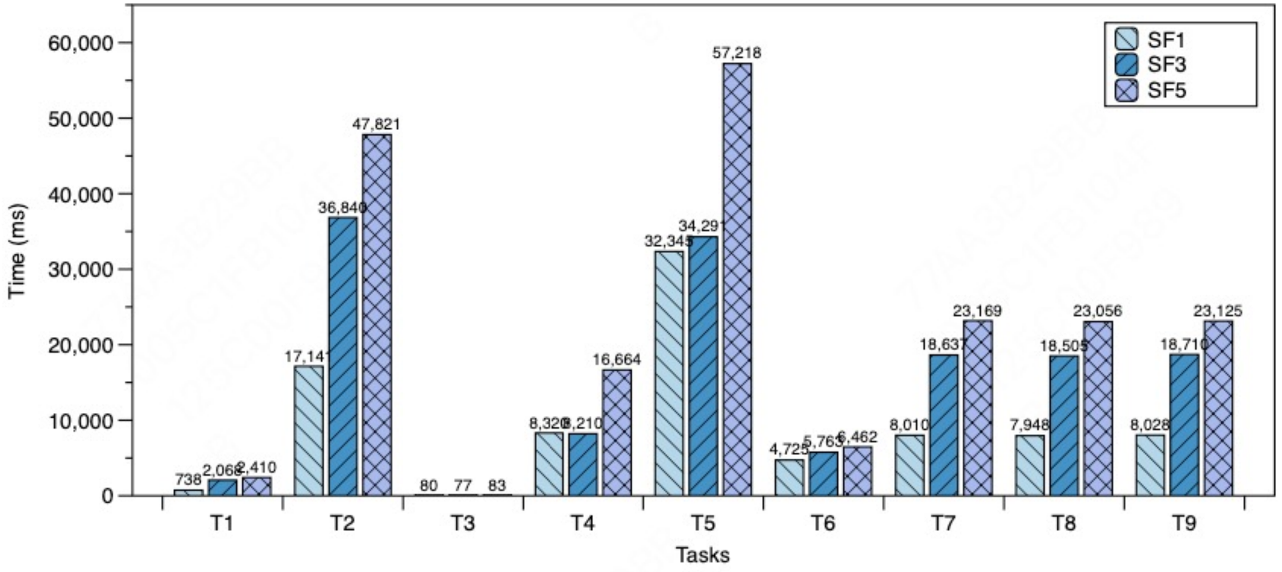


Fig.7. Elapse Time in Different Scales

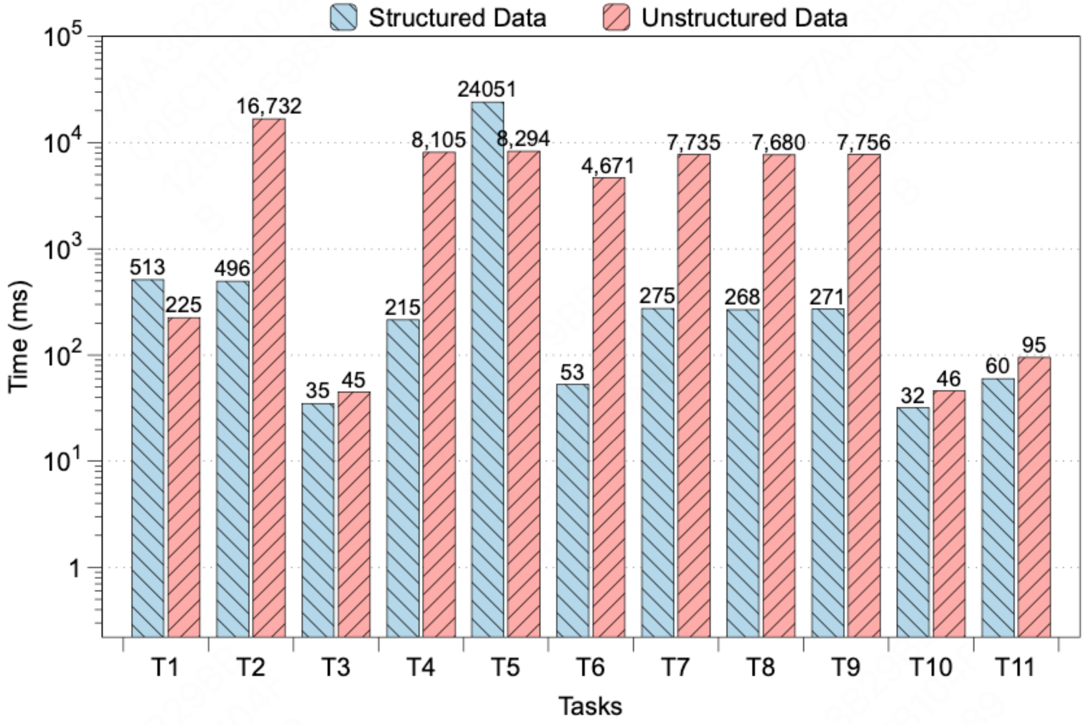
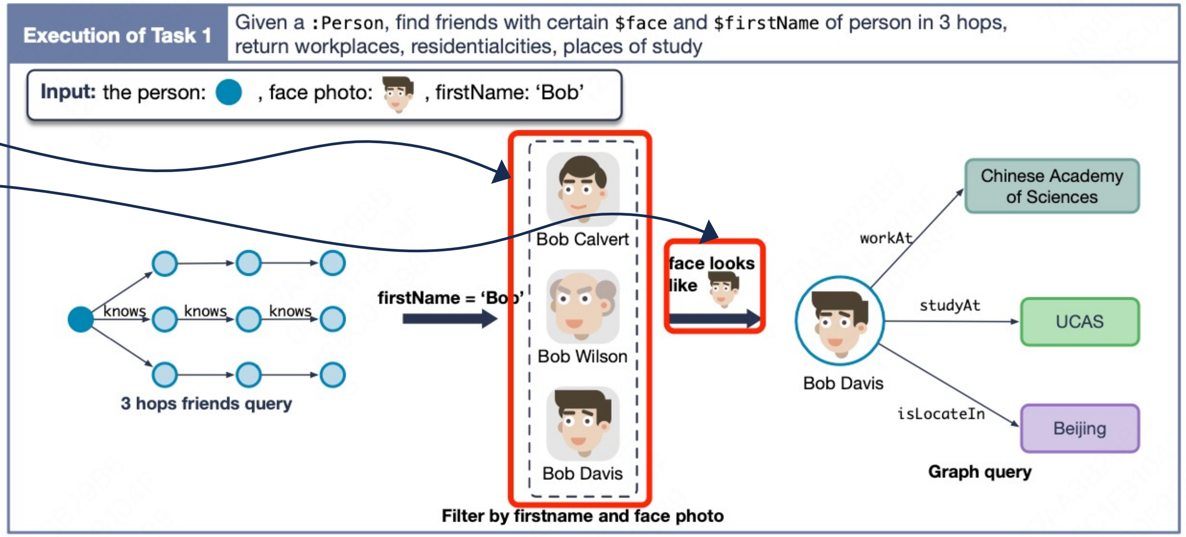


Fig.6. Processing Time for Structured Data and Unstructured Data

Experiment on Performance Improvement

Multiple-Step-Solution Latency

- Frequent connection establishment with OSS
- Frequent calls to HTTP request AI
- **Root cause**: Absence of Local Storage and Query Engine


Execution of Task 1
Given a :Person, find friends with certain $face and $firstName of person in 3 hops, return workplaces, residentialcities, places of study
Input: the person: ● , face photo: , firstName: 'Bob'
knows
knows
knows
3 hops friends query
firstName = 'Bob'
Bob Calvert
Bob Wilson
Bob Davis
Filter by firstname and face photo
face looks like
Bob Davis
workAt
studyAt
isLocateIn
Chinese Academy of Sciences
UCAS
Beijing
Graph query

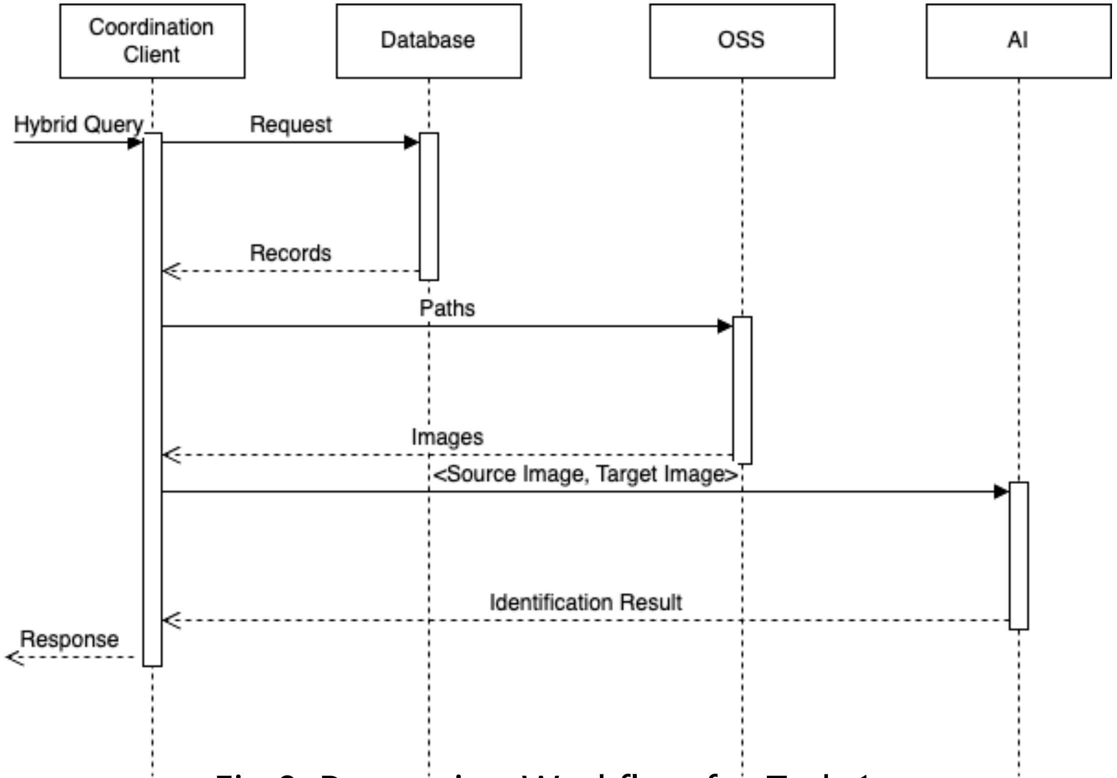


Fig.8. Processing Workflow for Task 1

In order to eliminate the latency, the experimental setup is as follows:

- Multimodal data is stored using a local file system
- Centralization of AI computing services and data storage services

Table 6. Improvement after optimization

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11
59%	53%	12%	93%	79%	30%	94%	98%	95%	7%	10%

$$Improvement\ Rate = (original\ time - improved\ time)/original\ time$$

Experiment Conclusion

- **Optimizing Hybrid Queries**:
  - Context: Absence of structured data indexing and caching.
  - Approach: Filter structured conditions first.
  - Impact: Substantially reduces search space for unstructured property and accelerates hybrid queries.
- **Structured Index vs. Unstructured Cache**:
  - Context: Short query-intensive tasks (Task10, Task11).
  - Observation: Performance gap not distinctly noticeable.
- **Challenges in Multimodal Querying**:
  - Traditional Approach: Multi-Step-Solution, which has latency.
  - Observation: Unstructured data query time exceeds structured data query time in most scenarios, which is a critical bottleneck in large-scale datasets.

Future Plans:

- Utilize AIGC (Artificial Intelligence for Generating Content) for generating higher quality and larger-scale datasets.
- Conduct experiments on real multimodal database to further validate the effectiveness of the proposed benchmark.