IEEE/ACM ICCAD
2025 INTERNATIONAL CONFERENCE ON COMPUTER-AIDED DESIGN
44th Edition

MM-GRADE: A Multi-Modal EDA Tool Documentation QA Framework Leveraging Retrieval Augmented Generation

Yuan Pu[1,2], Zhuolun He[1,2], Shutong Lin[1], Jiajun Qin[1], Xinyun Zhang[1], Hairuo Han[1], Haisheng Zheng[2], Yuqi Jiang[3], Cheng Zhuo[3], Qi Sun[3], David Pan[4], **Bei Yu**[1]

[1]The Chinese University of Hong Kong
[2]ChatEDA Tech [3]Zhejiang University
[4]University of Texas at Austin

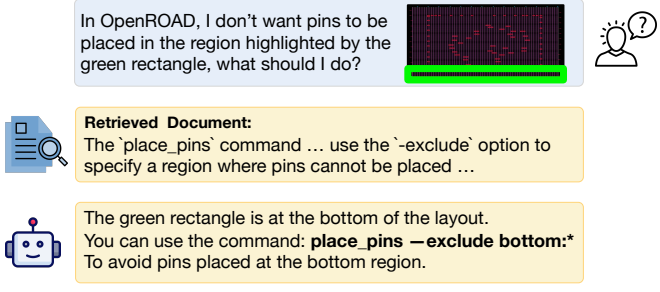
October 29, 2025

Outline

Introduction

EDA Tool Documentation QA

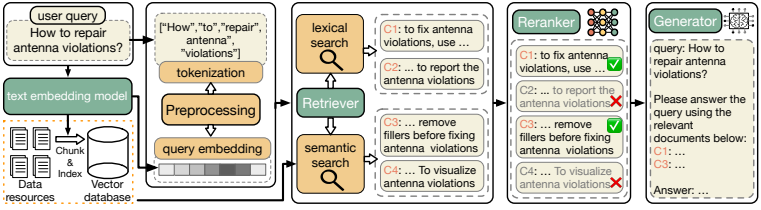
- EDA tool involves powerful but complex functionalities/documents → steep learning curve.
- Multi-Modal Automatic Question-Answering (QA) system:
  - Given the **user query**, the QA system generates **answer** referring to the relevant **documents**.
  - User query/document contain useful visual information:
    - GUI screenshots.
    - Design layouts.
    - Screenshot of error traceback.



A typical multi-modal query in OpenROAD.

RAG-based EDA QA Systems and Its Limitations

- Trend: Use LLMs and Retrieval Augmented Generation for EDA Tool QA.
- Existing works: RAG-EDA[1], etc.
- Stages:
  - Retriever: retriever the relevant documents to the user query.
  - Generator: usually LLM, generates the answer.
- Limitations: They are designed for **text-only** query, ignore **visual information** → fundamental gap in understanding the user's true intent.



Overall flow of RAG-EDA.

[1] **Yuan Pu et al. (2024). "Customized Retrieval Augmented Generation and Benchmarking for EDA Tool Documentation QA". In: *arXiv preprint*.**

General-Purpose Multi-Modal RAG Methods: Limitations

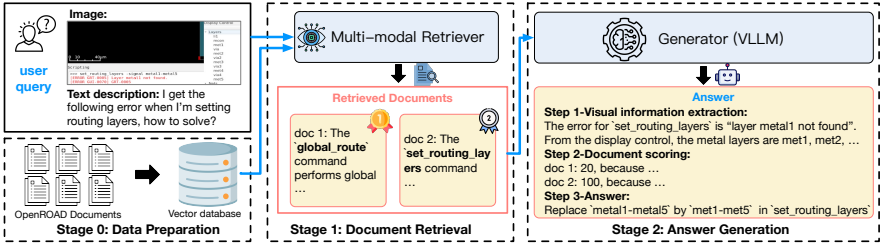
- Existing Multi-Modal RAG Methods: VisRAG[2], EchoSight[3], etc.
- Limitations of being applied to EDA tool documentation QA:
  - Retriever: General MM-RAG retrievers are trained on natural scenes, charts → struggle to comprehend the complex, specialized information in EDA visuals.
  - Generator: General Vision-Language Models (VLMs) lack EDA domain knowledge → *spurious understanding* of the user's query and poor quality answers

[2] **Shi Yu et al. (2024). "VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents".** In: *arXiv preprint*.

[3] **Yibin Yan and Weidi Xie (Nov. 2024). "EchoSight: Advancing Visual-Language Models with Wiki Knowledge".** In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 1538–1551.

Proposed Algorithm

Overall Flow



- Solution: MM-GRADE, a multi-modal, RAG-based EDA tool QA framework.
- Techniques:
  - Multi-modal Retriever: Bi-level negative mining strategy → contrastive learning.
  - Generator (vision large language model): customized 3-stage reasoning pipeline: Extract-Score-Answer (ESA) pipeline.
- Benchmark: ORD-MMBench, consisting of 120 QAs (from OpenROAD) with screenshots.

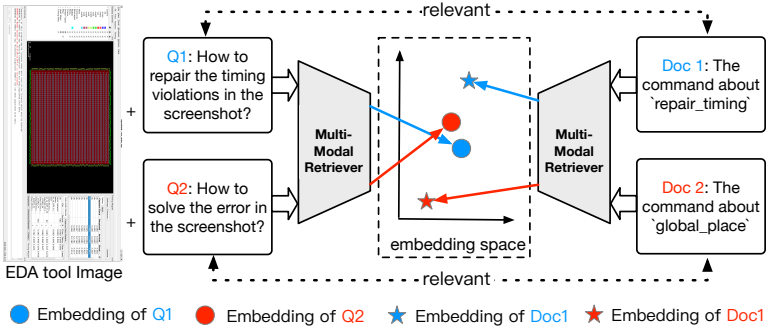
Existing Training Strategy of Multi-Modal Retriever

- Mechanism of multi-modal retriever for document retrieval:
  - Step 1: project the user query and all document chunks (multi-modality) into a unified embedding space.
  - Step 2: retrieve the most relevant document chunks by similarity matching.
- Training strategy: Contrastive learning with **document-level hard negative mining**:
  - Given a user query q.
  - positive samples: its relevant documents d_i^+
  - negative samples: its irrelevant documents d_i^-
- Training loss:

$$L_d^i = -\log \frac{e^{f(q_i^{\text{img}} \| q_i^{\text{txt}})^\top f(d_i^+)/\tau}}{e^{f(q_i^{\text{img}} \| q_i^{\text{txt}})^\top f(d_i^+)/\tau} + \sum_{j=1}^{n} e^{f(q_i^{\text{img}} \| q_i^{\text{txt}})^\top f(d_{i,j}^-)/\tau}}, \tag{1}$$

Limitations of Multi-Modal Retriever

- Observation: One EDA tool screenshot contains complex and diverse information → triggers multiple, distinct user questions.
- Consequence: Many different EDA tool user questions share one common image (screenshot).
- Limitation: Exiting general-purpose Multi-Modal retrievers focus on visual information → distinct user queries (with the same image) are projected closely in the embedding space → Low document retrieval accuracy.
- Example:



Solution of EDA-Customized Multi-Modal Retriever: Bi-level Hard Negative Mining

- Purpose: train the multi-modal retriever such that $\rightarrow$ distinct queries with the same image $\rightarrow$ projected **distinctly** in the embedding space $\rightarrow$ higher retrieval accuracy.
- Solution: contrastive learning with **query-level hard negative mining**:
  - Given a user query q.
  - positive samples: its relevant documents d_i^+
  - negative samples: distinct queries with **the same image** $\{(q_i^{\text{img}}, q_{i,k}^{\text{txt}}) | k \in [1, m] \wedge k \neq j\}$
- Training loss:

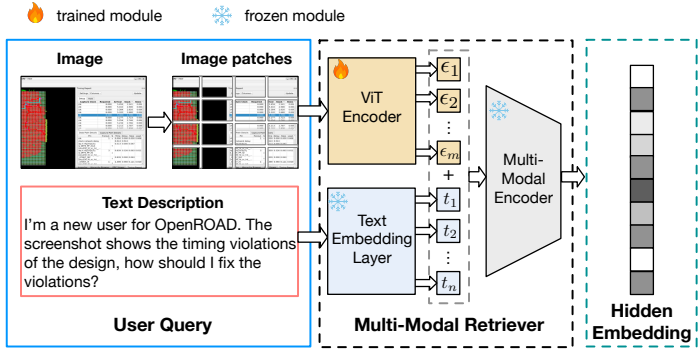
$$L_q^i = -\sum_{j=1}^{m} \log \frac{e^{f(q_i^{\text{img}} \| q_{i,j}^{\text{txt}})^T f(d_{i,j}^+)/\tau}}{e^{f(q_i^{\text{img}} \| q_{i,j}^{\text{txt}})^T f(d_{i,j}^+)/\tau} + \sum_{k \neq j} e^{f(q_i^{\text{img}} \| q_{i,j}^{\text{txt}})^T f(q_i^{\text{img}} \| q_{i,k}^{\text{txt}})/\tau}}. \tag{2}$$

- Bi-level hard negative mining (HNM): integrate document-level and query-level HNM:

$$L = \sum_{i=1}^{N} L_d^i + L_q^i. \tag{3}$$

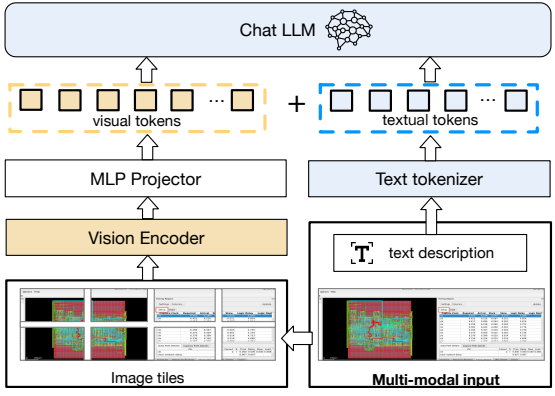
Multi-Modal Retriever: Proposed Model Architecture

- Components:
  - ViT encoder: encodes the visual input.
  - Text embedding layer: encodes the textual input
  - Multi-modal encoder: text-visual fusion for embedding generation.



Multi-Modal Generator: Model Architecture and Limitations

- Model architecture: based on InternVL2[4].
- Limitations when applied to EDA tool QA: lacks domain-specific inference logic.



[4]**Zhe Chen et al. (2024). "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks".** In: *Proc. CVPR*, pp. 24185–24198.

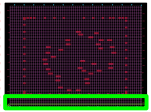
Solution of Multi-Modal Generator: ESA Pipeline

Extract-Score-Answer pipleline:

- Pipeline the answer into 3 stages:
  - Stage 1: Visual information extraction.
  - Stage 2: Document relevance scoring.
  - Stage 3: Answer
- Purpose: used for both training and inference.
- Scoring Criteria of Stage 2, between query q and document d:

Score	Description
100	d is perfectly relevant to q and can be used to directly answer q.
80	d is relevant to q and can partially answer q.
60	d is relevant to q but can not be directly used to answer q.
40	It is possible that d is relevant with q, but more information is required to verify the relevance.
20	d is basically not relevant with q.
0	d is totally irrelevant with q.

User visual query-image:



User visual query text description:
I don't want the pins to be placed on the region of the green rectangle in the screenshot, what should I do?

Retrieved Documents:
Doc 1: `place_pin`: the command to place a specific pin …

Doc 2: use the option `exclude` of `place_pins` …

Extraction-Scoring-Answer Strategy

Stage 1: Visual information extraction
Prompt: Extract the information from the image which can be used to understand and answer the question.
Answer: The pins highlighted in the green rectangle is the bottom region of the layout.

Stage 2: Document Relevance Scoring
Prompt: Now given the retrieved documents, analyse and score the relevance between each document and the visual query. The score ranges from 0 to 100, higher scores represent higher relevance.
Answer:
Doc 1: 20, because the query is about the command to place all pins, but `place_pin` command only places one pin.
Doc 2: 100, because the document introduces the usage of `place_pins` and the `exclude` flag.

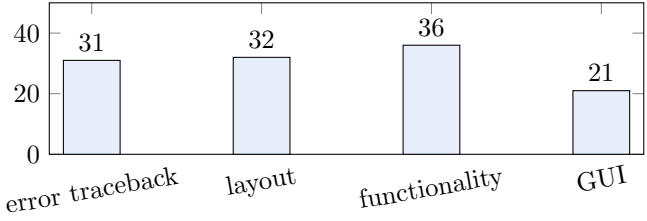
Stage 3: Answer
Prompt: answer this user query according the extracted visual information in stage 1 and the document relevance scoring results in stage 2.
Answer:
you can use the command: `place_pins -hor_layers <hor layers> -ver_layers <ver layers> -exclude bottom:"` to disable pins placed on the bottom region.

Prompt Example of the ESA Pipeline

Experimental Results

Benchmark: ORD-MMBench

- ORD-MMBench [5]: 120 high-quality multi-modal question-document-answer triplets based on the OpenROAD[6] document.
- Four categories of questions:



[5] https://github.com/lesliepy99/MM-GRADE-Benchmarks-ICCAD

[6] **Tutu Ajayi et al. (2019). "Toward an open-source digital flow: First learnings from the openroad project".** In: *Proc. DAC*, pp. 1–4.

Experimental Setting

- Multi-Modal Retriever:
  - Base model: bge-visualized-m3[7].
  - Training data: 1075 query-document corpus items generated by gpt-4o.
  - Finetuned by 3000 steps (batch size = 8) on 4 A100 80G GPUs.
- Multi-Modal Generator:
  - Base model: InternVL2-26B[8].
  - Training data: 2307 high-quality query-document-answer triplets based on OpenROAD[9].
  - Trained by 2 epochs (batch size = 1) on 8 A100 80G GPUs.

[7] https://huggingface.co/BAAI/bge-visualized/blob/main/Visualized_m3.pth

[8] https://huggingface.co/OpenGVLab/InternVL2-26B

[9] **Tutu Ajayi et al. (2019). "Toward an open-source digital flow: First learnings from the openroad project".** In: *Proc. DAC*, pp. 1–4.

Experimental Results: Retrieval Baselines

- Metric: recall@k, defined as the proportion of relevant documents that are retrieved among the top k results returned by the retriever/reranker.
- Baselines:
  - Text-only retriever: BGE-M3[10], RAG-EDA-retriever
  - Multi-modal retriever: DEDR[11], BLIP[12], VisRAG-retriever[13], EchoSight-retriever[14], bge-visualized-m3

[10] **Jianlv Chen et al. (2024).** ***BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*****.** **arXiv: 2402.03216 [cs.CL].**

[11] **Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani (2023). "A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering". In:** ***Proc. SIGIR*****, pp. 110–120.**

[12] **Junnan Li et al. (2022). "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation". In:** ***Proc. ICML*****. PMLR, pp. 12888–12900.**

[13] **Shi Yu et al. (2024). "VisRAG: Vision-based Retrieval-augmented Generation on Multi-modality Documents". In:** ***arXiv preprint*****.**

[14] **Yibin Yan and Weidi Xie (Nov. 2024). "EchoSight: Advancing Visual-Language Models with Wiki Knowledge". In:** ***Findings of the Association for Computational Linguistics: EMNLP 2024*****. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 1538–1551.**

Experimental Result: Multi-Modal Retrieval

- Ablation study: train MM-GRADE retriever (ours) with only query-level/document-level hard negative mining (HNM).
- MM-GRADE retriever with bilevel HNM achives the best performance.

Model type	recall@1	recall@2	recall@3	recall@4	recall@5
BGE-M3	0.590	0.694	0.731	0.761	0.776
RAG-EDA-retriever	0.530	0.612	0.679	0.724	0.739
DEDR	0.112	0.187	0.231	0.284	0.299
BLIP	0.127	0.209	0.276	0.299	0.313
VisRAG-retriever	0.025	0.042	0.050	0.083	0.108
EchoSight-retriever	0.383	0.492	0.583	0.625	0.667
VISTA: bge-visualized-m3	0.582	0.716	0.739	0.761	0.799
MM-GRADE-retriever w. d-level HNM	0.664	0.784	0.806	0.828	0.851
MM-GRADE-retriever w. q-level HNM	0.582	0.694	0.746	0.761	0.828
MM-GRADE retriever (ours)	**0.679**	**0.828**	**0.843**	**0.866**	**0.896**

Table: Performance of document retrieval on ORD-MMBench.

Experimental Result: Overall RAG Flow

- Metrics:
  - LLM-Score: gpt-4o to score each answer (from 0 to 100) referring to a scoring criteria [15].
  - BLEU and Rouge-L.
- RAG flow baselines:
  - Text-only: RAG-EDA.
  - Multi-modal: VisRAG, EchoSight, VISTA-RAG.

RAG Flow	ORD-MMBench▶error traceback			ORD-MMBench▶layout			ORD-MMBench▶functionality			ORD-MMBench▶GUI			ORD-MMBench▶all		
	LLM-Score	BLEU	ROUGE-L	LLM-Score	BLEU	ROUGE-L	LLM-Score	BLEU	ROUGE-L	LLM-Score	BLEU	ROUGE-L	LLM-Score	BLEU	ROUGE-L
VisRAG	41.129	0.080	0.223	12.500	0.098	0.242	18.750	0.111	0.247	4.762	0.024	0.162	20.417	0.084	0.225
EchoSight	51.613	0.087	0.241	54.688	0.165	0.300	62.500	0.191	0.322	64.286	0.091	0.214	57.917	0.139	0.276
RAG-EDA	22.419	0.064	0.187	61.719	0.156	0.276	59.722	0.190	0.332	51.190	0.107	0.267	49.125	0.134	0.268
VISTA-RAG	65.323	0.126	0.266	68.750	0.216	0.348	71.528	0.232	0.368	73.810	0.116	0.261	69.583	0.180	0.317
MM-GRADE w/o ESA	71.129	0.159	0.303	78.125	0.273	0.418	75.833	0.291	0.442	**90.476**	0.160	0.312	77.792	0.229	0.377
MM-GRADE (ours)	**82.258**	**0.244**	**0.391**	**80.469**	**0.307**	**0.468**	**80.000**	**0.294**	**0.450**	83.333	**0.178**	**0.323**	**81.292**	**0.264**	**0.417**

Table: Performance of the multi-modal RAG flows on ORD-MMBench.

[15] https://github.com/lesliepy99/MM-GRADE-Benchmarks-ICCAD

THANK YOU!