



DESIGN, AUTOMATION
AND TEST IN EUROPE

THE EUROPEAN EVENT FOR
ELECTRONIC SYSTEM DESIGN & TEST

20 – 22 APRIL 2026
VERONA, ITALY

PALAZZO DELLA GRAN GUARDIA



RATuner: Retrieval-Augmented VLSI Flow Design Parameter Tuning Framework

Peng Xu^{1,†}, Ziyang Yu^{1,†}, Yuan Pu¹, Xinyun Zhang¹, Donger Luo², Hao
Geng², Siyuan Xu³, Tsung-yi Ho¹, Bei Yu¹

¹CUHK ²ShanghaiTech University ³Huawei

April 22, 2026



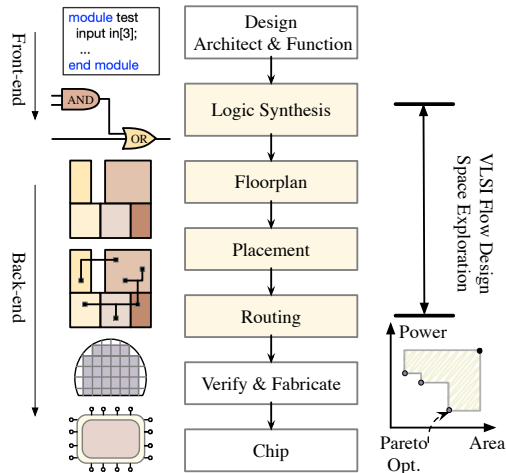
① Introduction

② Method

③ Experiments

Introduction

- Huge search space
 - $> 10^{70}$ configurations
- Expensive evaluations
 - Each flow run can take hours
- Strong inter-stage dependencies
 - Synthesis $\rightarrow$ Routing
- Domain knowledge in EDA manuals is unused by Bayesian Optimization(BO)



Sequential VLSI flow.

Motivation

- **Complex parameter semantics**: manual context is **critical** for Parameter Tuning.
- **Retriever + BO framework**: retrieves **exact documentation** instead of directly using LLMs as **surrogate model**.

Engineering Impact

- Provides **better feature engineering** for traditional BO.
- Avoids **LLM hallucination** when directly using LLMs for config selection.

Design Stage:

Floorplan

Design Parameter:

-coreMarginsBy

Related context:

Specifies whether the core margins are calculated using the core-to-I/O boundary or the core-to-die boundary.
Default: io

Design Choices:

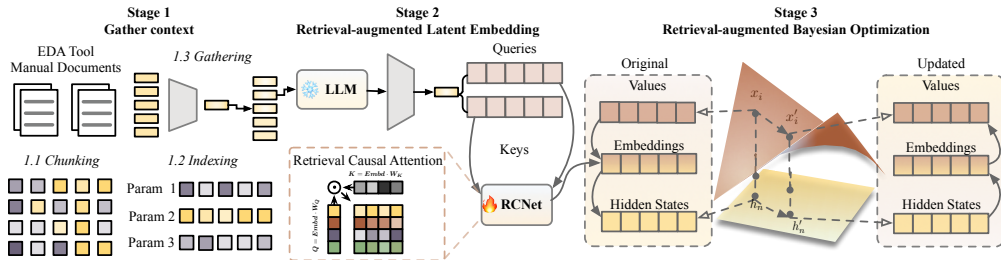
{io | die}

Related context:

Specifies the floorplan dimensions by size; or by die, I/O, or core coordinates. The floorPlan command initializes the floorplan and calls the add_tracks command to create new routing tracks.

Example documentation context.

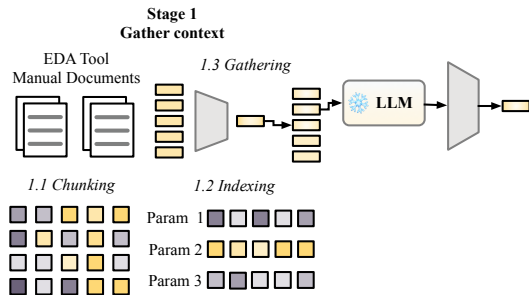
Method



Three-stage pipeline: retrieval, causal embedding, and latent BO.

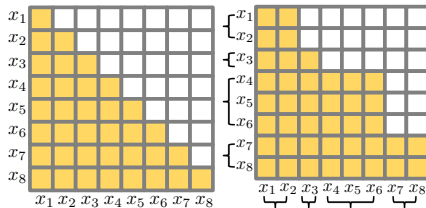
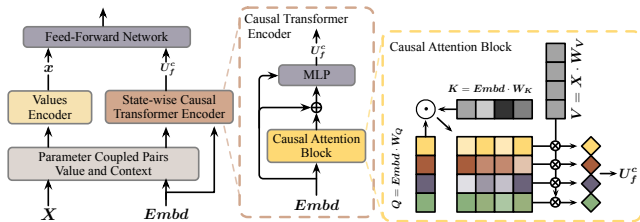
The stages work collaboratively: retrieval provides domain context, causal embedding models inter-stage dependencies, and latent BO optimizes efficiently in the learned space.

- 1 **Chunk** EDA manuals into uniform text segments.
- 2 **Index** parameter names via rule-based parsing.
- 3 **Gather** matched chunks as parameter contexts.
- 4 **C-RAG^a** captures EDA-specific terminology.



^aRAG-2024iccad-pu.

2. RCNet: Causality-Aware Embedding



- Retrieval context vectors + parameter values.
- Stage-wise causal mask enforces inter-stage dependencies.
- Produces compact latent embeddings for BO.

Physical Design Irreversibility

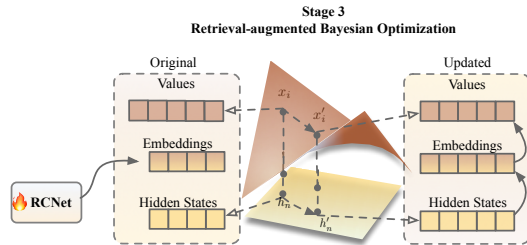
RCNet's causal mask ensures later-stage embeddings do not influence earlier-stage embeddings, modeling this one-way dependency.

Standard BO

- 1 Surrogate Model (GP)
- 2 Acquisition Function
- 3 Evaluate

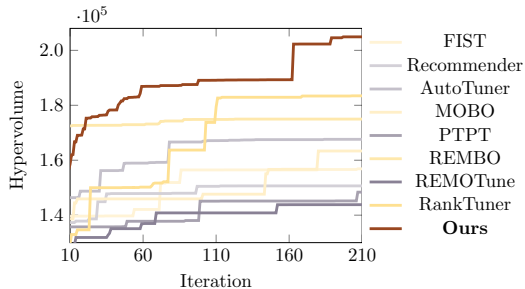
RABO (Our Method)

- 1 **RA**: Retrieve documentation context
- 2 **RCNet**: Build causal latent embedding
- 3 Surrogate Model (DGP) in latent space
- 4 Acquisition Function (Thompson sampling)
- 5 Evaluate & Update latent space



- **RA** injects **domain knowledge** before optimization.
- **Latent space** updated **iteratively** with evaluated QoR.

Experiments



Hypervolume convergence.

- **Up to 33% HV improvement** vs. best baseline.
- **Faster convergence**: RATuner reaches higher HV with fewer iterations.
- **Better sample efficiency**: achieves superior QoR with similar or fewer expensive flow evaluations.
- **Consistent gains** across all three benchmarks.

Table: HV improvements on RISC_V_32I.

Method	HV	MPI1 (%)	MPI2 (%)	MAI (%)
FIST	1.57	3.16	3.90	5.47
REMOTune	1.75	4.38	6.27	7.45
RankTuner	1.84	13.64	5.04	5.12
RATuner (Ours)	2.05	15.54	8.48	7.44

- **+11.4% HV** vs. RankTuner (best baseline).
- **+17.1% HV** vs. REMOTune.
- **+30.6% HV** vs. FIST.

Table: HV improvements on RISC_V_Rocket.

Method	HV	MPI1 (%)	MPI2 (%)	MAI (%)
FIST	1.47	12.38	-0.51	-1.01
REMOTune	1.61	16.11	1.57	-1.31
RankTuner	1.67	13.00	5.09	-0.96
RATuner (Ours)	1.73	13.85	6.69	-0.96

- **+3.6% HV** vs. RankTuner (best baseline).
- **+7.5% HV** vs. REMOTune.
- **+17.7% HV** vs. FIST.

THANK YOU!