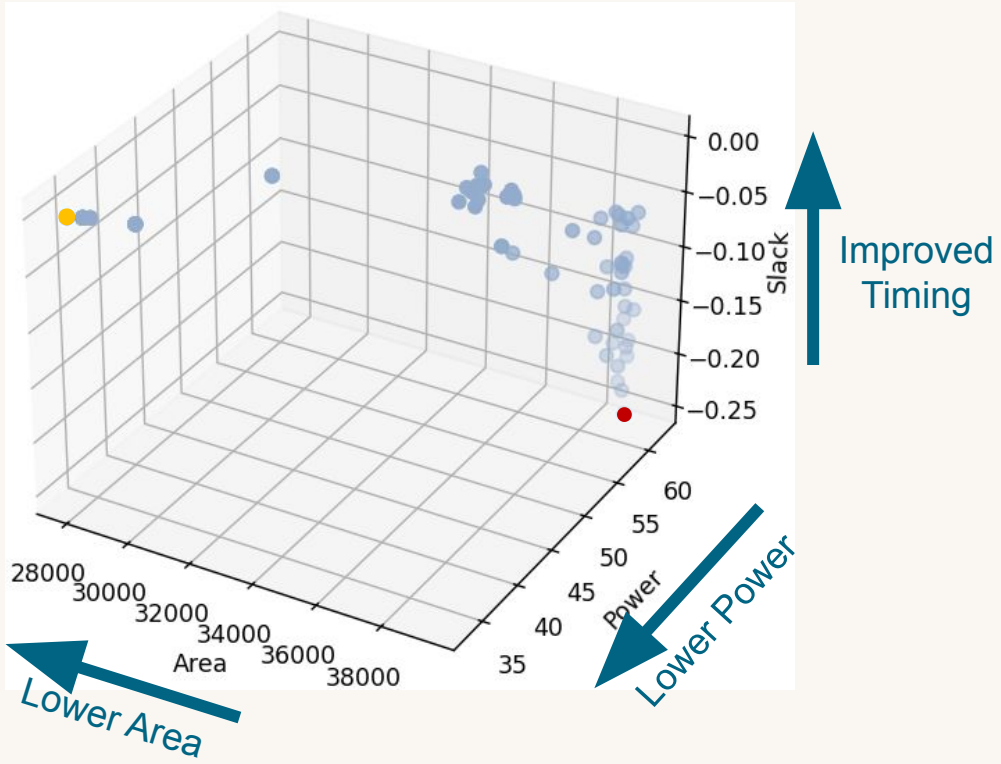
AstroTune: AST-Assisted LLM Retrieval for Cross-Stage Design Flow Parameter Tuner

Runzhi Wang[1], Jingyu Pan[2], Yiran Chen[2], **Jiang Hu[1]**

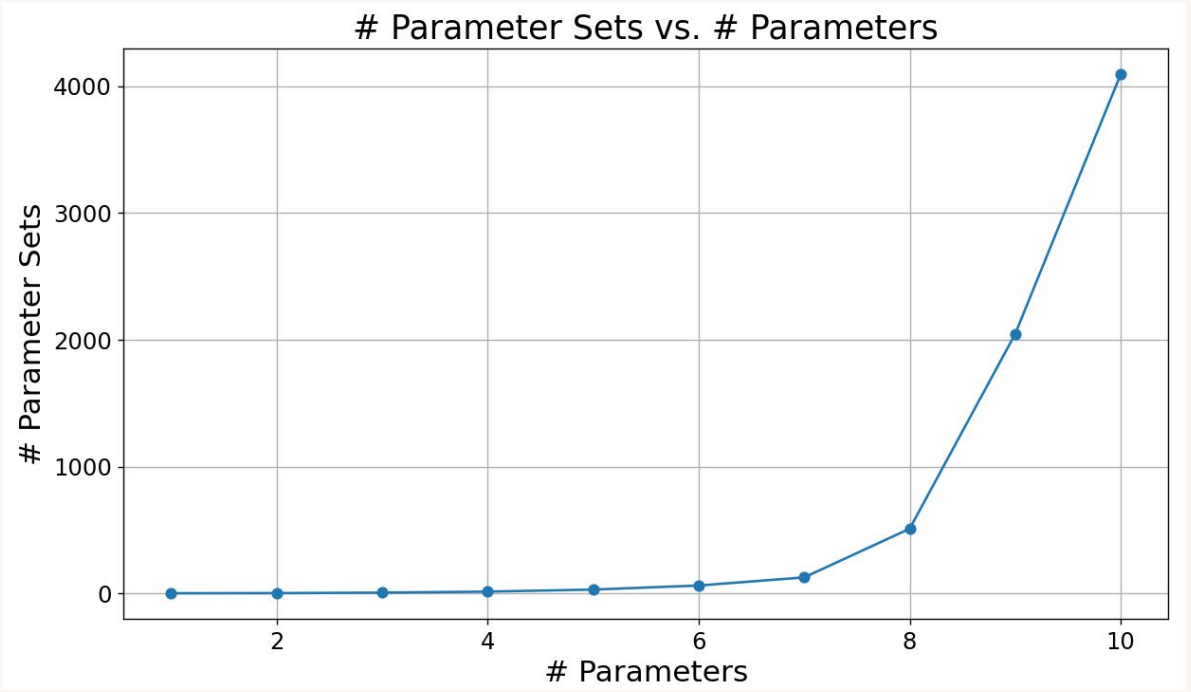
Texas A&M University[1], Duke University[2]

2026 International Symposium on Physical Design

Why Parameter Tuning Matters?



Parameter change → Huge PPA variation



Labor intensive! ***Time consuming!***

Impractical to enumerate

Existing Solutions

Method	HDL Semantic Awareness	RTL Structure Awareness	Stage-by-Stage Tuning	Automatic
Manual Design	✓ High	✓ High	✓ High	✗
Bayesian Optimization	✗	✗	✗	✓
SynTunSys[1]	✗	✗	✗	✓
FIST[2]	✗	✓ Basic	✗	✓
FlowTuner [3]	✓ Basic	✗	✓ Basic	✓
ORFS-agent[4]	✓ Basic	✗	✓ Basic	✓
CROP [5]	✓ High	✓ Basic	✗	✓

[1] [Ziegler+, *Design, Automation & Test in Europe Conference & Exhibition (DATE)*'2016]

[2] [Xie+, *Asia and South Pacific Design Automation Conference (ASP-DAC)*'2020]

[3] [Liang+, *International Conference on Computer Aided Design (ICCAD)*'2021]

[4] [Ghose+, *International Symposium on Machine Learning for CAD (MLCAD)*'2025]

[5] [Pan+, *International Conference on Computer Aided Design (ICCAD)*'2025]

Existing Solutions

We need **domain knowledge-based** approach that adapts **faster** without **manual intervention**!

	HDL Semantic Awareness	RTL Structure Awareness	Stage-by-Stage Tuning	Automatic
	✓ High	✓ High	✓ High	✗
	✗	✗	✗	✓
	✗	✗	✗	✓
	✗	✓ Basic	✗	✓
	✓ Basic	✗	✓ Basic	✓
	✗	✗	✓ Basic	✓
	✓ High	✓ Basic	✗	✓

[1] [Ziegler+, *Design, Automation & Test in Europe Conference & Exhibition (DATE)*'2016]

[2] [Xie+, *Asia and South Pacific Design Automation Conference (ASP-DAC)*'2020]

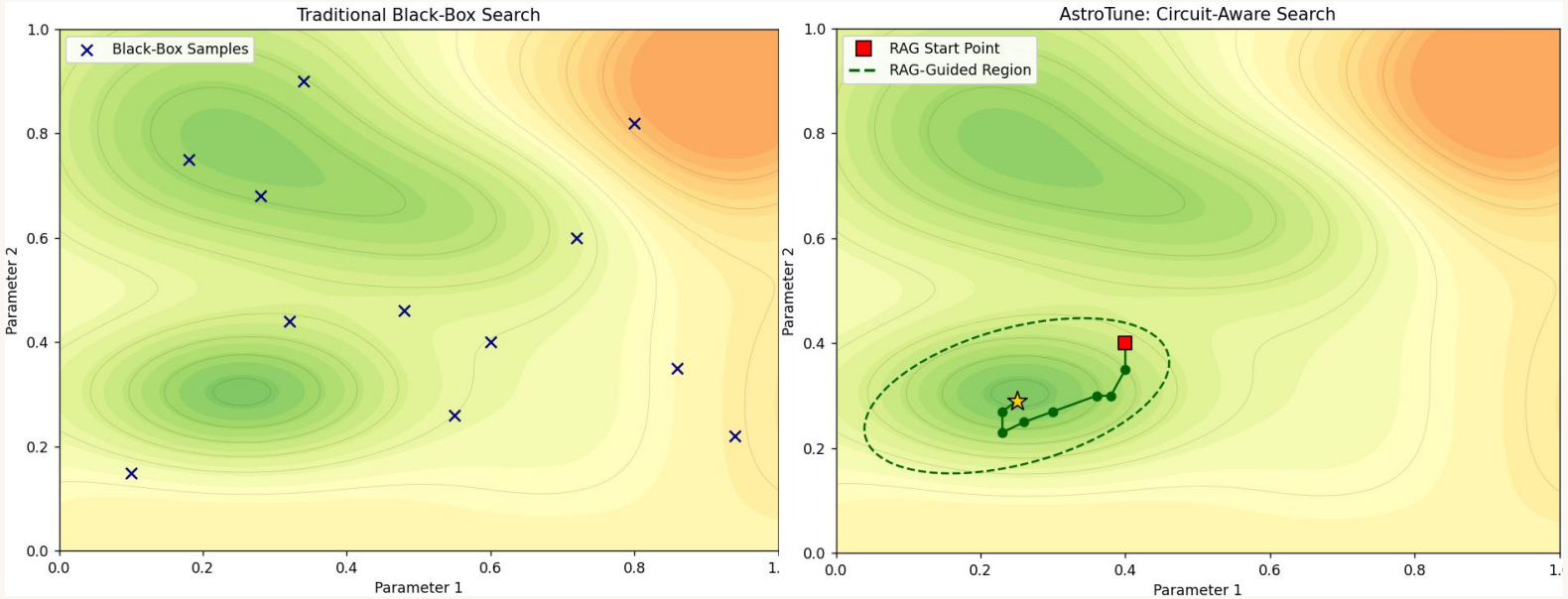
[3] [Liang+, *International Conference on Computer Aided Design (ICCAD)*'2021]

[4] [Ghose+, *International Symposium on Machine Learning for CAD (MLCAD)*'2025]

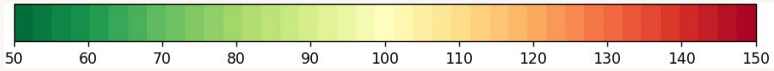
[5] [Pan+, *International Conference on Computer Aided Design (ICCAD)*'2025]

Key Idea

1. Retrieval-Based Warm Start
2. Tournament-based Pruning
3. Stage-by-Stage Tuning

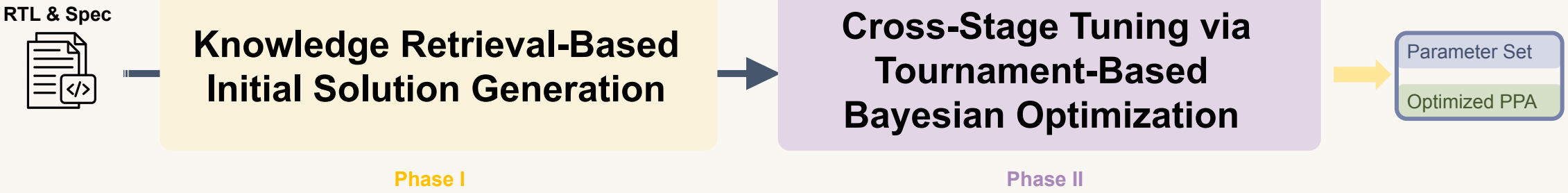


Fewer trials but better results

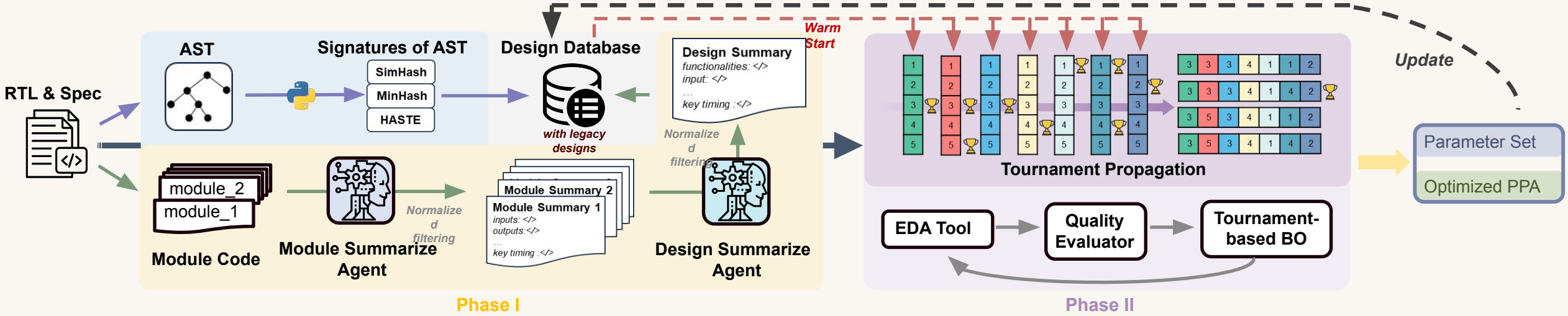


QoR score is the smaller the better

Overall Workflow: AstroTune



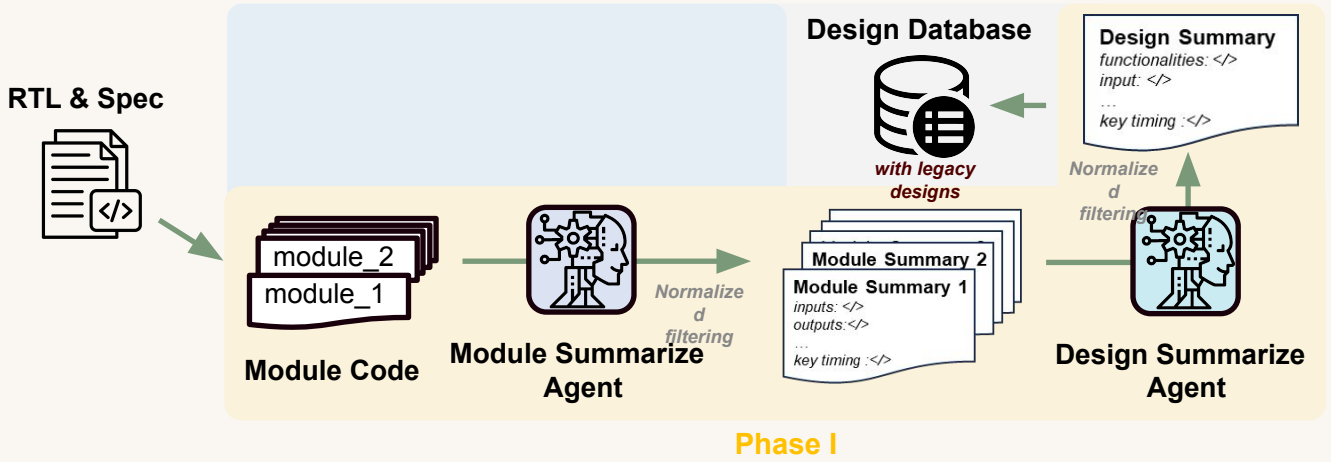
Overall Workflow: AstroTune



Accumulate experience as human engineers

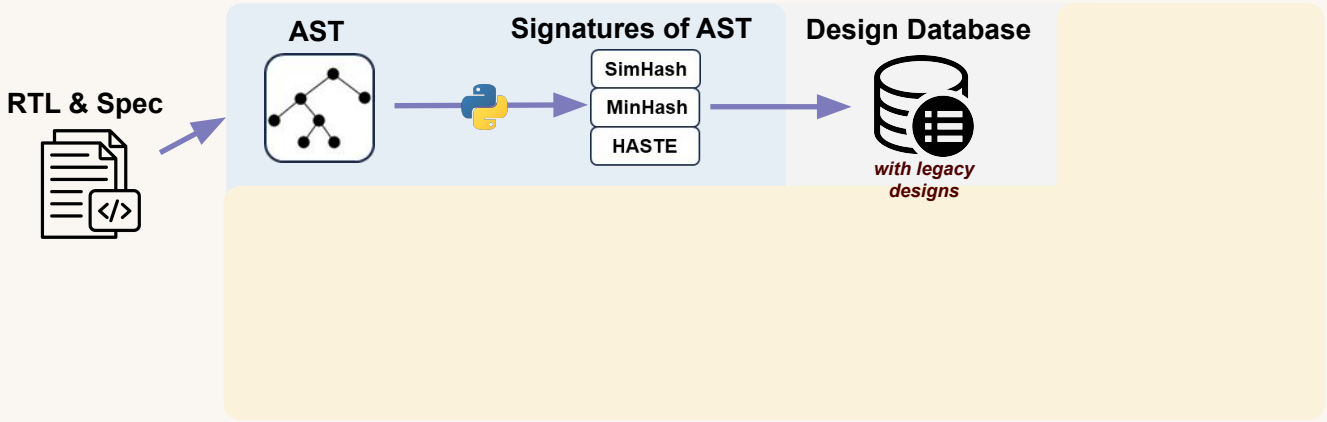
Phase I: Knowledge Retrieval-Based Initial Solution Generation

Semantic Information



Phase I: Knowledge Retrieval-Based Initial Solution Generation

Structural Information



Abstract syntax tree(AST) → Signatures

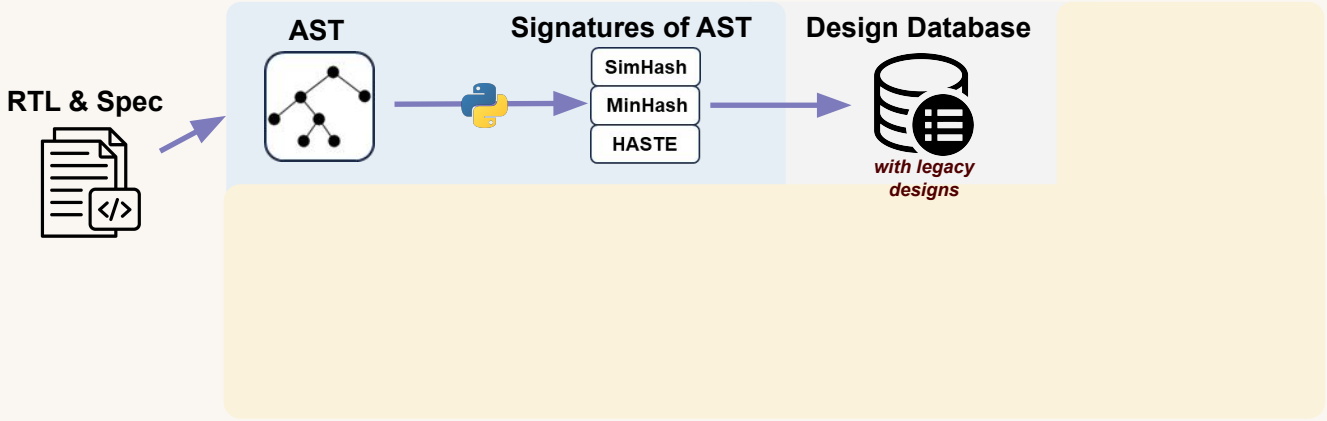
Signature from SimHash[6]	Signature from MinHash [7]	Signature from Hashed AST Embedding (HASTE)
Global vector sketch	Set-overlap estimator	Token-level vector embedding
Loses fine details	*Set-only view*	*Too sensitive*

[6] [Charikar +, *Symposium on Theory of Computing (STOC)*'2002]

[7] [Wu+, *IEEE Transactions on Knowledge and Data Engineering*.'2020]

Phase I: Knowledge Retrieval-Based Initial Solution Generation

Structural Information



Signature from SimHash[6]

$$\boldsymbol{\psi}[i] = \mathbf{1}\left\{\sum_{\boldsymbol{\tau}\in\tilde{T}_R} w_\tau(2\boldsymbol{\tau}[i]-1)\geq 0\right\}, i=0,\ldots,b-1$$

$$S_{sim}(\hat{T}_{R_i},\hat{T}_{R_j}) = 1-\frac{\eta\left(\boldsymbol{\psi}(\hat{T}_{R_i}),\boldsymbol{\psi}(\hat{T}_{R_j})\right)}{b}$$

Signature from MinHash [7]

$$m_i\left(\hat{T}_R\right) = \min_{\boldsymbol{\tau}\in\hat{T}_R} H(\boldsymbol{\tau};s_i), i=1,\ldots,k$$

$$S_{min}(\hat{T}_{R_i},\hat{T}_{R_j}) = \frac{1}{k}\sum_{i=1}^{k}\mathbf{1}\left\{m_i(\hat{T}_{R_i})=m_i(\hat{T}_{R_j})\right\}$$

Signature from Hashed AST Embedding (HASTE)

$$\hat{\mathbf{v}}(\hat{T}_R) = \frac{\sum_{\boldsymbol{\tau}\in\hat{T}_R} w_{\boldsymbol{\tau}}\,\sigma(\boldsymbol{\tau})\,\mathbf{e}_{h(\boldsymbol{\tau})}}{\left\|\sum_{\boldsymbol{\tau}\in\hat{T}_R} w_{\boldsymbol{\tau}}\sigma(\boldsymbol{\tau})\,\mathbf{e}_{h(\boldsymbol{\tau})}\right\|_2} \in \mathbb{R}^d$$

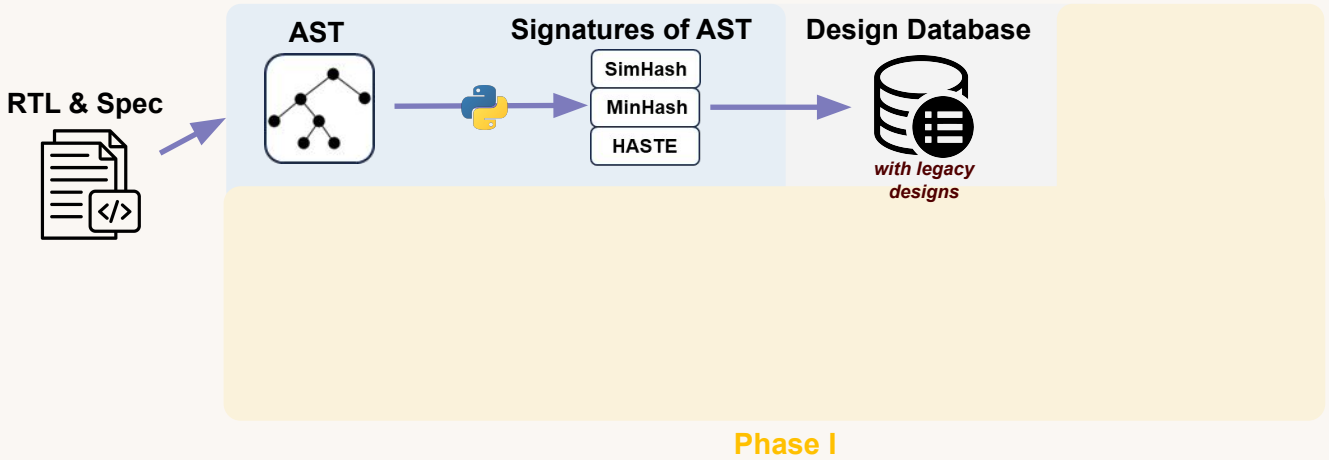
$$S_{min}(\hat{T}_{R_i},\hat{T}_{R_j}) = \frac{1}{k}\sum_{i=1}^{k}\mathbf{1}\left\{m_i(\hat{T}_{R_i})=m_i(\hat{T}_{R_j})\right\}$$

[6] [Charikar +, *Symposium on Theory of Computing (STOC)*'2002]

[7] [Wu+, *IEEE Transactions on Knowledge and Data Engineering*.'2020]

Phase I: Knowledge Retrieval-Based Initial Solution Generation

Structural Information

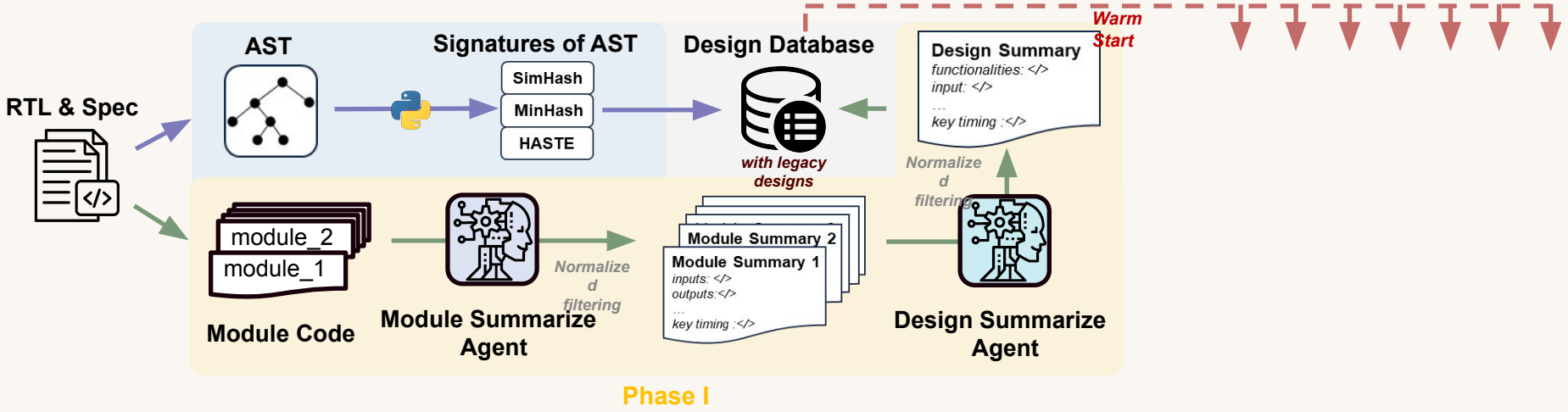


Overall Similarity

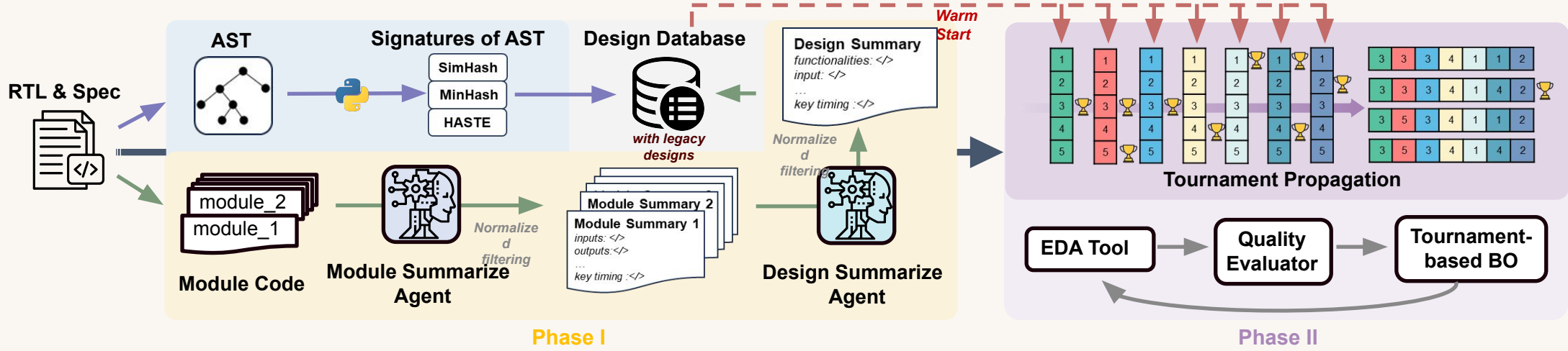
$$S_{overall}(\hat{T}_{R_i}, \hat{T}_{R_j}) = \sum_{\ell \in \{\text{sim}, \text{min}, \text{haste}, \text{sem}\}} \gamma_\ell \, S_\ell(\hat{T}_{R_i}, \hat{T}_{R_j})$$

$\hat{T}_{R_i}, \hat{T}_{R_j}$: token sets of designs being compared.

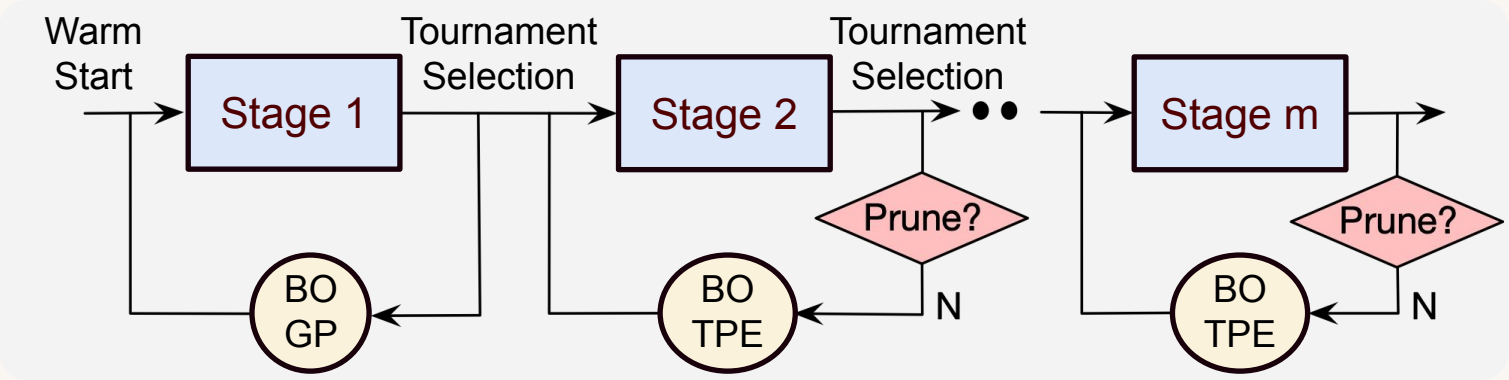
Phase I: Knowledge Retrieval-Based Initial Solution Generation



Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization



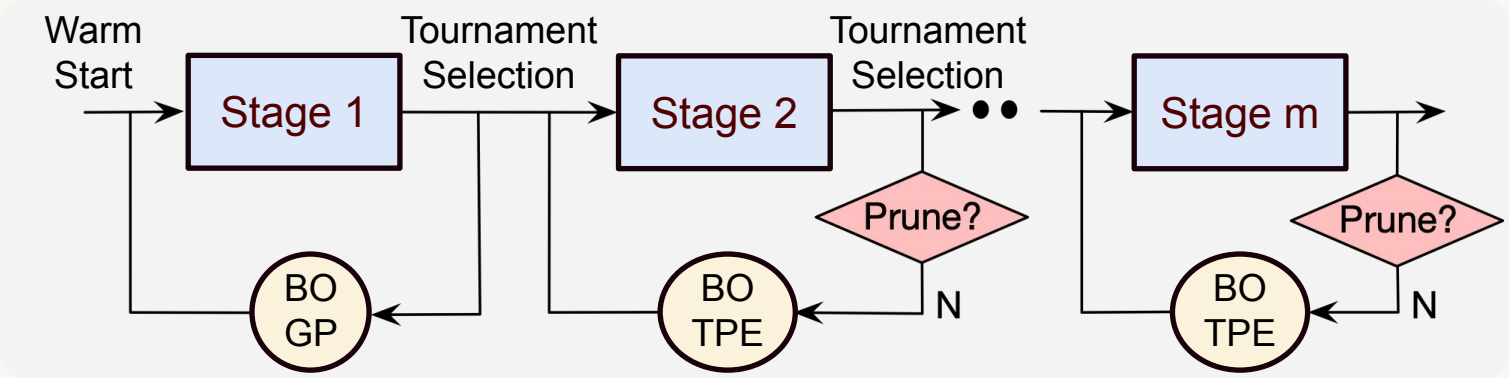
Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization



Promote a solution to next stage, when no better candidate emerges over several attempts.

GP(Gaussian Process) : Uses Gaussian Processes as a surrogate model to predict the objective function.
TPE(Tree-structured Parzen Estimator) : Uses Probability Density Estimation as a surrogate function to guide sampling.

Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization



Promote a solution to next stage, when no better candidate emerges over several attempts.

- Prune using a threshold set by the stage's first solution.
- Use per-stage surrogate models.
- Early-stop is triggered when multiple non-improving attempts exceed the minimum attempts.

Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization

QoR score

Observation:

Small logic synthesis time violations are often fixed during layout

Action:

The smaller the violation, the lighter the penalty.

Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization

QoR score

Logic Synthesis

$$r = \frac{\max(0, -\mathcal{W} - \lambda_1 \cdot \mathcal{T})}{\max(\mathcal{W}_{ref}, \lambda_2 \cdot \mathcal{T})}.$$ *Segmented time violation penalty*

$$\mathcal{Q}_{logic} = \alpha_{\mathcal{A}} \cdot \frac{\mathcal{A}}{\mathcal{A}_{ref}} + \alpha_{\mathcal{P}} \cdot \frac{\mathcal{P}}{\mathcal{P}_{ref}} + \lambda_3 \cdot \min(r, \beta) + \lambda_4 \cdot \max(r - \beta, 0)$$

Layout Synthesis

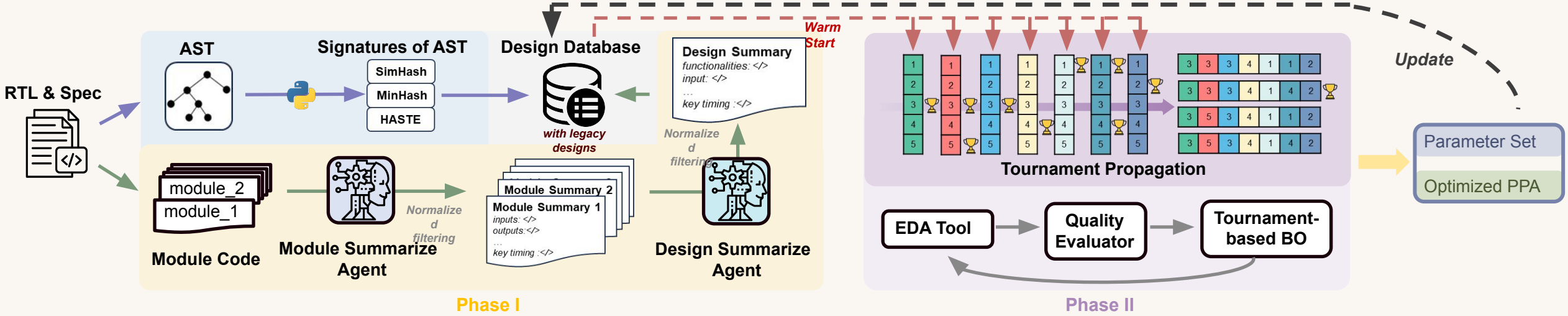
$$\mathcal{Q}_{layout} = \begin{cases} \kappa + \alpha_{\mathcal{W}} \cdot |\mathcal{W}|, & \mathcal{W} < 0 \\ \alpha_{\mathcal{A}} \cdot \frac{\mathcal{A}}{\mathcal{A}_{ref}} + \alpha_{\mathcal{P}} \cdot \frac{\mathcal{P}}{\mathcal{P}_{ref}}, & \mathcal{W} \geq 0 \end{cases}$$

$\mathcal{P}$: Power $\mathcal{W}$: Performance $\mathcal{A}$: Area

$\mathcal{T}$: Clock period β : Breakpoint(a multiple of the clock period)

κ : Penalty

Phase II: Cross-Stage Tuning via Tournament-Based Bayesian Optimization



Enriching database along with usage

Experiment Setup

- Logic synthesis is performed with Synopsys Design Compiler using a 45nm standard cell library.
- Physical design steps are carried out on Cadence Innovus.

- A Linux x86_64 machine equipped with an AMD Ryzen Threadripper PRO 5955WX 16-core processor (32threads), 256 GB memory, and an NVIDIA GeForce RTX 4090 GPU(24 GB)

- Total of 33 parameters distributed across 7 stages of the flow: logic synthesis, placement, placement optimization, Clock Tree Synthesis(CTS), CTS optimization, routing, and routing optimization.

Datasets

- Testcases are significantly larger and diversity.

- Designs in database and testing sets are separated.

Work	Num of designs	Number of cells			
		Min	Median	Mean	Max
ORFS-agent [4]	3	12055	17265	21060	33861
FlowTuner [3]	5	3467	35314	38668	75912
CROP (Public) [5]	14	3	60	493	4377
AstroTune testcases*	20	437	14874	56978	726859

*AstroTune testcases contains all testcases of both [3] and [4] .

[3] [Liang+, *International Conference on Computer Aided Design (ICCAD)*'2021]

[4] [Ghose+, *International Symposium on Machine Learning for CAD (MLCAD)*'2025]

[5] [Pan+, *International Conference on Computer Aided Design (ICCAD)*'2025]

Comparison with Different Approaches

Average Results	AstroTune	CROP [5]	BO	Optuna [8]	LLM Search	FlowTuner [3]	Logfile-aware LLM Search [4]
Area(μm^2)	126639 (1×)	131165 (1.04×)	130674 (1.03×)	132255 (1.04×)	133229 (1.05×)	131399 (1.04×)	130075 (1.03×)
Power(mW)	100.769 (1×)	115.235 (1.14×)	114.601 (1.14×)	116.585 (1.16×)	117.626 (1.17×)	115.429 (1.15×)	114.834 (1.14×)
WS (ns)	0.0177	-0.0567	-0.0513	-0.0655	-0.0688	-0.0231	-0.0588
QoR Score	90.831 (1×)	104.588 (1.15×)	104.480 (1.15×)	104.696 (1.15×)	105.328 (1.16×)	103.585 (1.14×)	104.973 (1.16×)
Run Time(min)	633 (1×)	1509 (2.38×)	1351 (2.13×)	1619 (2.56×)	1538 (2.43×)	1923 (3.03×)	1537 (2.43×)

Area, power, QoR score are the smaller the better
Worst slack(WS) is the larger the better

AstroTune achieved the best across all metrics

[3] [Liang+, *International Conference on Computer Aided Design (ICCAD)*'2021]

[4] [Ghose+, *International Symposium on Machine Learning for CAD (MLCAD)*'2025]

[5] [Pan+, *International Conference on Computer Aided Design (ICCAD)*'2025]

[8] [Akiba+, *ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD)*'2019]

Detailed results of all testcases

AstroTune winning in most test cases

Ablation Study: Retrieval Quality

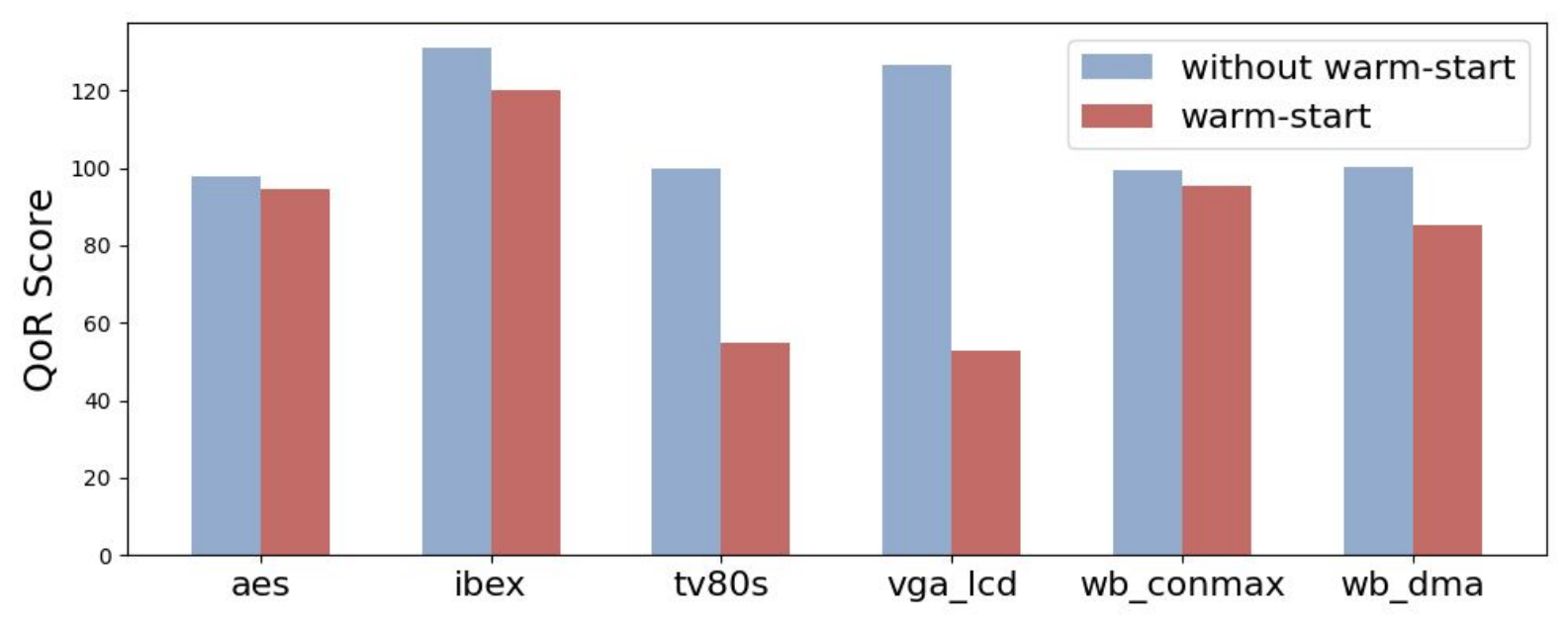
Design	Semantic from LLM		SimHash		MinHash		HASTE		AstroTune	
	Original	Altered	Original	Altered	Original	Altered	Original	Altered	Original	Altered
HR@3(%)	100	12.5	87.5	87.5	33.3	33.3	33.3	33.3	100	100
HR@1(%)	95.8	0	87.5	87.5	33.3	33.3	33.3	33.3	100	95.8

Altered: Anonymize module name and remove comments.

HR@k(%) denotes the percentage of cases where thecorrect design is retrieved within the top k candidates.

AstroTune utilizes structural information for highly accurate retrieval.

Ablation Study: Impact of Warm-Start.



Warm-start can get better QoR of PPA

Conclusion

- Structure-aware retrieval
- Robust knowledge transfer
- Tournament-based cross-stage tuning
- Per-stage surrogate models
- Better QoR with fewer trials

THANK YOU