Automating Hardware Design and Verification from Architectural Papers via a Neural-Symbolic Graph Framework

Anonymous ACL submission

Abstract

Reproducing hardware architectures from academic papers is critical yet challenging due to the semantic gap between unstructured text and formal logic. The abstract nature of documentation often hinders the distillation of high-level architectural intent into rigorous implementations. We propose **ArchCraft**, a multi-agent framework that treats hardware generation as a neural-symbolic comprehension task. By formalizing blueprints into hierarchical graphs and requirements into symbolic specifications, ArchCraft ensures that the generated Register-Transfer Level (RTL) is strictly grounded in the paper's logic. We utilize downstream physical synthesis as a structural validation metric to ensure architectural alignment and implementation feasibility rather than mere code similarity. Furthermore, we introduce **ArchSynthBench**, the first benchmark featuring 64 project-level circuits for evaluating automated hardware reproduction. Experimental results demonstrate that ArchCraft significantly outperforms baselines in architectural integrity. By leveraging physical implementation as a rigorous feasibility check, we demonstrate that ArchCraft effectively translates high-level textual intent into physically viable and architecturally faithful hardware logic.

1 Introduction

"This paper on a novel neural network accelerator is excellent," your hardware director mused, "but its implementation is closed-source. We must reproduce it to build our next-generation products upon it." This demand highlights a fundamental challenge, Figure 1: how can we design an automated system to seamlessly convert a 30-page unstructured PDF into rigorous, synthesizable code? (Qian et al., 2023; Seo et al., 2025; Zhao et al., 2025c,b; Zhang et al., 2023a) While a human expert typically deconstructs this task into a multi-step mental model of hierarchical mapping and functional specification prior to RTL and testbench (TB) translation, current AI paradigms struggle to replicate this specialized workflow, but they often fail to grasp project-level architectural blueprints due to training data constraints. Meanwhile, existing Agent Approaches often manifest as flat or less structured approaches that rely on waveforms or TBs as primary inputs, thereby sidestepping the core challenge of true generation from scratch: translating high-level textual specifications into functional hardware logic.

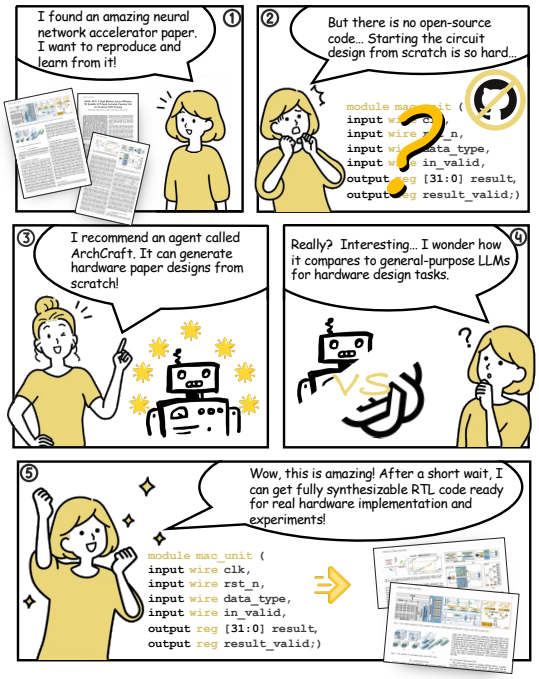

Figure 1: A real-world example. ArchCraft assists in automating the process of transforming hardware design concepts into fully synthesizable RTL code, reducing manual intervention and accelerating the design process.

We thus explore ArchCraft, a training-free, structured agentic framework that maintains workflow autonomy by inheriting the expert's mental model. Our core premise is: *effective hardware generation requires replacing general-purpose reasoning*

with a domain-specific, formalized intermediate representation. At its heart, the Graphing Agent constructs the architectural blueprint, a process involving scope definition, node/link identification, attribute planning, and global constraint integration. Subsequently, the Symbolization stage translates these graphical attributes into a rigorous symbolic representation. This symbolic blueprint serves as a precise intermediate language, constraining the LLM to adhere to hardware logic without fine-tuning. Building on this foundation, dedicated Coding and Evaluating agents generate the RTL and TB, followed by comprehensive functional verification. To substantiate the fidelity of the reconstructed designs, we extend the verification to the physical level. By integrating downstream logic synthesis, we utilize physical Power, Performance, and Area (PPA) characteristics as a rigorous metric to validate that the generated logic aligns with the architectural scale and intent of the source paper, ensuring the designs meet the rigors of hardware implementation.

Furthermore, we introduce ArchSynthBench, the first benchmark for reconstructing hardware from closed-source accelerator papers, featuring 64 project-level circuits with approximately 650 functional blocks. A hierarchical evaluation across eight dimensions demonstrates that ArchCraft significantly outperforms established agent frameworks. Specifically, ArchCraft achieves a paper-level fidelity of 86.18, a 24.8% improvement over the state-of-the-art, while reaching a code-level completeness of 81.48, effectively establishing a new performance upper bound for project-level hardware reconstruction. These results, corroborated by physical synthesis trends, validate the framework's capability to ground abstract documentation into complex, synthesizable hardware in a zero-shot manner.

- We propose **ArchCraft**, a novel neural-symbolic graph-based framework that enables the complex task of transforming unstructured academic papers into synthesizable, project-level hardware implementations.
- Our proposed symbolic representation of functions and interfaces bridges the gap between RTL documentation and code.
- We construct a benchmark, **ArchSynthBench**, to rigorously evaluate the replication accuracy and implementation quality of hardware synthesis.
- We further conduct physical-level verification to substantiate the framework's effectiveness in grounding high-level architectural intent.

2 Related Work

2.1 LLMs for Research

LLM-agents are increasingly central to scientific discovery, from ideation to execution. Frameworks decompose research into autonomous modules for tasks like mining, experimentation, and writing (Schmidgall and Moor, 2025; Schmidgall et al., 2025; Ghafarollahi and Buehler, 2025; Baek et al., 2024; Cui et al., 2025; Zhao et al., 2025a), enhancing efficiency and scalability. For reproducibility, AutoReproduce (Zhao et al., 2025c) uses a paper lineage algorithm to extract implicit knowledge, automating AI experiment reproduction with demonstrated superior performance. Collectively, these works show the potential of multi-agent LLM frameworks to transform the scientific process.

2.2 LLMs for Hardware Design

LLM application in hardware design is growing. Foundational datasets underpin this progress, such as PyraNet (Nadimi et al., 2024), CodeV (Zhao et al., 2024), and MG-Verilog (Zheng et al., 2024). Methodologies include direct generation and agent-based frameworks. Direct generation includes Chip-Chat (Blocklove et al., 2023) for interactive co-design, and structured prompting methods like HiVeGen (Tang et al., 2024), ROME (Nakkab et al., 2024), SynC-LLM (Liu et al., 2025b) and VeriMind (Nadimi et al., 2025) (using chain-of-thought) to improve correctness and precision. Agent-based frameworks include VerilogCoder (Ho et al., 2025), which integrates planning, checking, and simulation, and RTLSquad (Wang et al., 2025), a collaborative agent architecture for integrated design, implementation, and verification, advancing hardware automation.

3 ArchCraft

In this section, we introduce ArchCraft. Unlike previous approaches, fine-tuning or circuit-level agents, our ArchCraft embeds a Mental Model. To ensure structural integrity, we employ formal mappings to precisely define the information boundaries and functional scopes of each specialized agent. These formulations act as a rigorous logical scaffold, preventing semantic drift and "hallucinations" by strictly constraining input-output dependencies across the project-level pipeline. The gener-

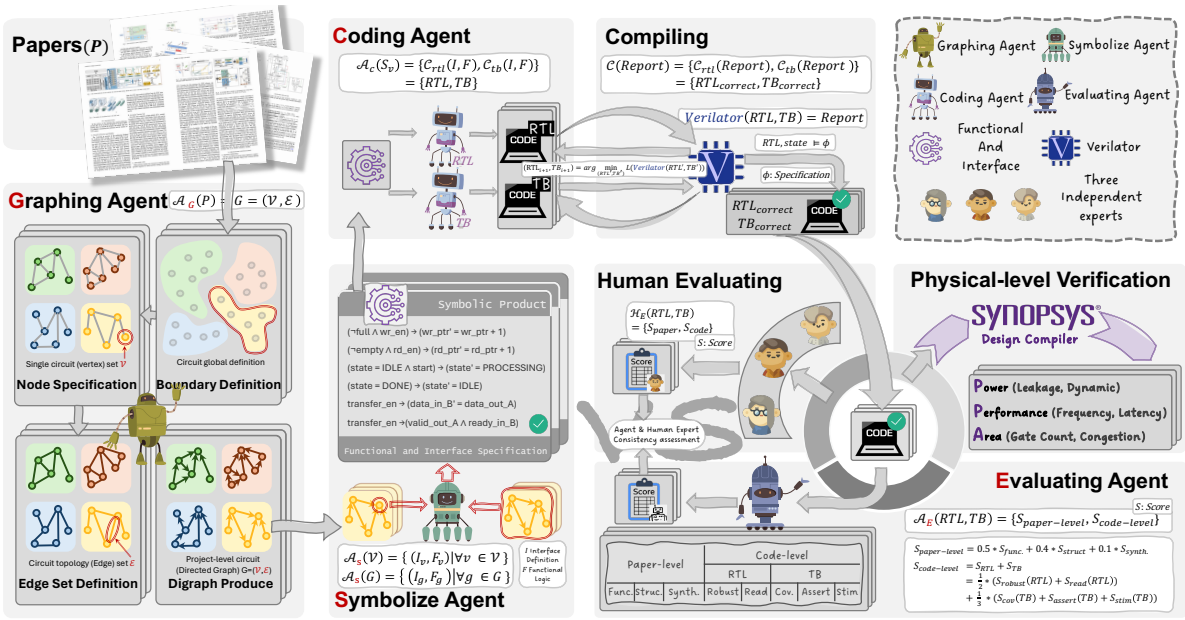

Figure 2: Overview of our ArchCraft framework with four phases. (1) Graphing Agent parses the paper into a formal architectural knowledge graph. (2) Symbolize Agent converts the graph nodes into a rigorous symbolic blueprint. (3) Coding Agent decoupledly generates synthesizable RTL and TBs from the symbolic blueprint. (4) Automatic execution of compilation and checking of grammar and logical issues, collects error reports, holds errors and related codes accountable, and collaboratively corrects them within a continuous feedback loop. Finally, both LLM and human experts evaluate the generated code to assess the overall effectiveness of the system. The generated RTL code is then synthesized into a hardware implementation for physical-level evaluation.

ated code supports comprehensive evaluation from functional to physical implementation. Prompts for the framework are in Appendix B.

3.1 Input Pre-processing and Modality Analysis

We utilize papers from the domain of architecture, particularly those related to ASICs, as input. In contrast to internal industrial documents, academic papers provide rigorous, publicly accessible documentation. We employ MinerU (Niu et al., 2025) to convert PDFs into structured text as our primary input. Our preliminary ablation studies (Appendix C) indicate that incorporating visual modalities yields negligible gains, as the dense textual descriptions in high-quality papers—such as logic analysis and functional specifications—sufficiently define the core architectural logic. However, these papers remain abstract regarding low-level circuit implementation details. A more detailed comparison is provided in Appendix Table 8. This neural-symbolic framework is therefore proposed to bridge the gap between abstract textual intent and concrete hardware implementation.

3.2 Graphing Phase

The Graphing phase, orchestrated by the Graphing agent $\mathcal{A}_{\mathcal{G}}$, transforms an academic paper P into a hierarchical architectural graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ representing hardware modules and their interconnections. This transformation resolves entity ambiguity in unstructured text through a four-step sequence:

Scope and Theme. First, the agent establishes the high-level context by identifying the core hardware accelerator or SoC architecture. This initial distillation defines the boundaries of the implementation scope S, such as the specific sub-module hierarchy, and the primary technical themes T, including sparsity-aware mechanisms or quantization-specific logic. Formally, this process can be described as:

$$\mathcal{A}_{\mathcal{G}_1} : P \mapsto (S, T) \tag{1}$$

where $\mathcal{A}_{\mathcal{G}_1}$ denotes the agent responsible for scope establishment.

Nodes and Static Links. Guided by the established scope S, the agent employs a recursive entity extraction strategy to identify hardware modules. Nodes $\mathcal{V}$ are instantiated based on hierarchical keywords like Processing Elements or Controllers,

while static edges $\mathcal{E}_{\text{static}}$ are established to represent structural containment or physical layout adjacencies. This process can be represented as a mapping:

$$\mathcal{A}_{\mathcal{G},2} : (P, S) \mapsto (\mathcal{V}, \mathcal{E}_{\text{static}}) \tag{2}$$

where $\mathcal{A}_{\mathcal{G}_2}$ denotes the agent responsible for nodes and static links construction.

Direction and Internal Nodes. With the node set $\mathcal{V}$ defined, the agent parses functional descriptions to determine dataflow orientation. Directed edges $\mathcal{E}_{\text{dir}}$ are constructed by identifying source-to-sink relationships within the text, such as direct buffer-to-multiplier pathways. Simultaneously, the agent assigns internal properties Φ, encompassing critical attributes like bit-widths and storage capacities, to each node $v \in \mathcal{V}$. This process is represented by:

$$\mathcal{A}_{\mathcal{G}_3}(P, \mathcal{V}) \mapsto (\mathcal{E}_{\text{dir}}, \Phi) \tag{3}$$

in which $\mathcal{A}_{\mathcal{G}_3}$ is the agent for edge generation.

Global Constraints. Finally, the agent extracts system-wide requirements C that enforce architectural consistency. These include clocking schemes, reset polarities, and standardized bus protocols, which are essential for ensuring cross-module synchronization. This extraction task is formalized as the following mapping:

$$\mathcal{A}_{\mathcal{G}_4}(P) \mapsto C \tag{4}$$

Here, $\mathcal{A}_{\mathcal{G}_4}$ denotes the agent responsible for identifying and extracting the set of global constraints C from the entire paper P.

The resulting graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ serves as the formal architectural blueprint. These high-level modules are subsequently decomposed into specific registers and IO ports during the Symbolization Phase to enable precise RTL generation.

3.3 Symbolization Phase

Following the construction of $\mathcal{G}$, the Symbolization phase, executed by $\mathcal{A}_{\mathcal{S}}$, translates abstract node attributes into a rigorous Symbolic Intermediate Representation (SIR). This SIR acts as a semantic bridge, converting natural language logic into precise hardware-aware rules. This stage functions as a rigorous semantic validation layer to ensure architectural fidelity; any fundamental misinterpretations from the Graphing stage typically emerge as logical inconsistencies within the symbolic constraints. Consequently, ArchCraft can identify and rectify structural discrepancies before the final implementation phase. As shown in Figure 2, the process operates at two levels:

Graph-level Symbolization. The agent first defines the top-level interface $I_{\mathcal{G}}$ (e.g., chip-level IO) and the global finite state machine (FSM) or control flow $F_{\mathcal{G}}$ that orchestrates the entire graph $\mathcal{G}$:

$$\mathcal{A}_{\mathcal{S}} : \mathcal{G} \mapsto (I_{\mathcal{G}}, F_{\mathcal{G}}), \qquad \text{where } \mathcal{G} = (\mathcal{V}, \mathcal{E}). \tag{5}$$

For intermediate subsystems $\mathfrak{G} \subseteq \mathcal{P}(\mathcal{G})$, the agent generates composite interface protocols that define how sub-module groups interact:

$$\mathcal{A}_{\mathcal{S}}(\mathfrak{G}) = \{(I_g, F_g) \mid g \in \mathfrak{G}\}. \tag{6}$$

Node-level Symbolization. For every leaf node $v \in \mathcal{V}$, $\mathcal{A}_{\mathcal{S}}$ produces a precise interface definition I_v (specifying signal names, widths, and handshake protocols like Valid/Ready) and a functional specification F_v, expressing logic via state transition rules or Boolean equations:

$$\mathcal{A}_{\mathcal{S}} : \mathcal{V} \to \mathcal{I} \times \mathcal{F}, \qquad v \mapsto (I_v, F_v), \tag{7}$$

and the image of the whole node set is

$$\mathcal{A}_{\mathcal{S}}(\mathcal{V}) \;=\; \{(I_v, F_v) \mid v \in \mathcal{V}\}. \tag{8}$$

By integrating these specifications, we obtain a comprehensive and multi-level representation of the project-level circuit design. The complete specification set $\{(I, F)\}$ effectively establishes a hierarchical constraint network that spans from top-level system coordination down to individual module-level logic. Formally, the project specification is defined by the union of specifications across all architectural layers:

$$\{(I, F)\} = \mathcal{A}_{\mathcal{S}}(\mathcal{V}) \;\cup\; \mathcal{A}_{\mathcal{S}}(\mathfrak{G}) \;\cup\; \{(I_{\mathcal{G}}, F_{\mathcal{G}})\} \tag{9}$$

Appendix D details the transition from text to symbolic and synthesizable RTL.

3.4 Coding and Compiling Phase

This phase transforms the interface I and functional specification F into synthesizable RTL code and its corresponding TB.

3.4.1 Coding

The Coding Agent, $\mathcal{A}_{\mathcal{C}}$ takes the symbolic interface I and functional logic F for each module as its input and generates two distinct code artifacts: the RTL implementation $\mathcal{C}_{rtl}$ and the TB $\mathcal{C}_{tb}$.

$$\begin{aligned} \mathcal{A}_{\mathcal{C}}(I, F) &= (\mathcal{C}_{rtl}(I, F), \mathcal{C}_{tb}(I, F)) \\ &= (RTL, TB) \end{aligned} \tag{10}$$

To prevent biased convergence, RTL and testbenches are generated independently. While both remain strictly grounded in the same formal specification to ensure alignment, we employ distinct prompting strategies and specialized agents for each task. This decoupling ensures the testbench serves as an independent functional auditor of the hardware logic.

3.4.2 Compiling and Internal Feedback Loop

The generated code artifacts $\{RTL, TB\}$ are not assumed to be correct on the first attempt. They immediately enter a rigorous verification and refinement loop, as illustrated in Figure 2.

Verification: The code is passed to an external simulator (e.g., Verilator) for compilation and execution, which produces a simulation $Report$.

$$Verilator(RTL, TB) = Report \quad (11)$$

Internal Reflection: The $Report$ is then fed back into a corrector function $\mathcal{C}_\mathcal{F}(Report)$. This function analyzes the report to identify discrepancies and determines if the current code is correct $\{RTL_{correct}, TB_{correct}\}$.

$$\begin{aligned} \mathcal{C}_\mathcal{F}(Report) &= (\mathcal{C}_{rtl}(Report), \mathcal{C}_{tb}(Report)) \\ &\stackrel{?}{=} (RTL_{correct}, TB_{correct}) \end{aligned} \quad (12)$$

Iterative Refinement: If errors are detected, e.g., $Report$ indicates failure, the framework activates an internal feedback loop. This iterative process, defined as $i+1$, seeks to find a corrected set of code (RTL', TB') that minimizes the error loss function L based on the output of the Verilator as:

$$\arg \min_{(RTL', TB')} L(Verilator(RTL', TB')) \quad (13)$$

This feedback cycle: Coding $\rightarrow$ Verilator $\rightarrow$ Report $\rightarrow$ Refinement repeats, continuously checking the RTL's state (Φ) against the original specification. The loop ends only when the code passes all the checks, resulting in verified artifacts $\{RTL_{correct}, TB_{correct}\}$.

4 ArchSynthBench

This section further introduces ArchSynthBench from three perspectives: the composition of the dataset (Section 4.1), the metrics and scoring methodologies for both machine and human expert evaluations (Section 4.2), and the physical implementation methods for the circuits (Section 4.3).

4.1 Data Source

ArchSynthBench consists of 64 distinct, closed-source hardware papers spanning 64 project-level circuits, and around **650** circuit blocks, which VerilogEval(Liu et al., 2023) has only 156 circuit blocks. This represents a significant expansion in project-level complexity over module-centric benchmarks like VerilogEval. Beyond its scale, ArchSynthBench introduces comprehensive criteria for evaluating agent workflows. The full paper list, along with detailed paper's theme, is available in Appendix Table 12 and Figure 5.

4.2 Evaluating Phase

Following recent LLM-as-judge approaches (Gu et al., 2024; Chen et al., 2024b), to quantitatively assess the quality of the generated artifacts, we introduce a comprehensive, two-tiered evaluation system orchestrated by an Evaluating Agent $\mathcal{A}_{\mathcal{E}}$, specific scoring criteria are set to measure the codes, detailed in Appendix F. This phase takes the verified RTL and TB as inputs and computes two distinct scores S, in Figure 2: $S_{paper-level}$ and $S_{code-level}$.

Paper-level Score. The $S_{\text{paper-level}}$ score measures the fidelity of the generated code to the original academic paper. It assesses how well the implementation reproduces the concepts described in the source text. This score is defined as a weighted sum of three sub-metrics: Functionality $S_{\text{Func.}}$, Structure $S_{\text{Struct.}}$, and Synthesizability $S_{\text{Synth.}}$, as follow:

$$S_{\text{paper-level}} = 0.5 \cdot S_{\text{func.}} + 0.4 \cdot S_{\text{struct.}} + 0.1 \cdot S_{\text{synth.}} \quad (14)$$

Detailed definitions for these sub-metrics are available in Appendix Table 9.

Code-level Score. The $S_{\text{code-level}}$ score assesses the intrinsic quality and robustness of the generated code itself, independent of the source paper. It is a weighted sum of the RTL Score S_{RTL}, the TB Score S_{TB} and the execution rate of the sub-circuits within each project-level circuit, with weights of 0.4, 0.4, and 0.2. The S_{RTL} is the average of the design's Robustness $S_{\text{Robust.}}$ and Readability $S_{\text{Read.}}$. The S_{TB} is the average of the verification environment's Coverage $S_{\text{Cov.}}$, Assertions $S_{\text{Assert.}}$, and Stimulus $S_{\text{Stim.}}$. Detailed evaluation rubrics are provided in Appendix Table9, 10.

Furthermore, the project-level execution rate is derived from the loop results (Section 3.4.2). These machine-generated scores are then used in conjunction with human expert scores for consistency

assessment, ensuring the automated metrics align with human judgment on design quality.

4.3 Physical-level Verification

Beyond functional correctness, we evaluate the feasibility of the generated hardware through an industry-standard synthesis flow. In this study, runnable implementation specifically refers to the RTL's capacity to achieve timing closure and successful logical synthesis.

As illustrated in Figure 2, the functionally correct RTL code is fed into an industry-standard logic synthesis tool, which synthesizes the RTL into a gate-level netlist, from which we extract the essential PPA metrics:

Power: We analyze the estimated power consumption, including both static Leakage power and Dynamic power dissipated during operation.

Performance: We evaluate the timing characteristics, primarily the operational Latency.

Area: We measure the total design footprint, reported as the logical Gate Count, and assess the routing difficulty via the Congestion report.

5 Experiments

5.1 Experiment Settings

Baseline. We propose the novel task of generating hardware RTL code from academic hardware papers. Given the absence of agentic workflow baselines specialized for this task, we compare our framework directly against LLMs. These LLMs include OpenAI o3-mini (OpenAI, 2025), Gemini 2.0 Flash (Google DeepMind, 2025), and Qwen3-Coder-480B-A35B-Instruct-FP8 (Team, 2025), all of which generate RTL code in a single step from the raw paper input, devoid of any intermediate workflow or agentic collaboration, thus effectively acting as unstructured generators.

For evaluating the broader efficacy of our framework, we also benchmark its performance against existing work in the domain of automated software generation, notably ChatDev (Qian et al., 2023), PaperCoder (Seo et al., 2025), and VerilogCoder (Ho et al., 2025) as an EDA-native baseline. All aforementioned baselines, along with our proposed ArchCraft framework, are evaluated using our custom ArchSynthBench benchmark. Crucially, the RTL and testbench code generated by all baselines were evaluated using identical standards and the same o3-mini evaluation agent applied to ArchCraft on ArchSynthBench, ensuring a fair and consistent comparison. To reduce potential evaluator bias, we further conducted cross-verification using heterogeneous models Qwen3-32B (Yang et al., 2025a) as independent scorers. Our analysis showed a high degree of correlation between this independent assessment and our primary o3-mini evaluator.

Hardware Implementation Environment. We conduct comprehensive hardware synthesis experiments using Synopsys DC, a widely adopted ASIC synthesis tool. All experiments are performed using Synopsys DC at 200MHz on SIMC 28nm technology on a high-performance server equipped with Intel Xeon Gold CPUs and 256 GB RAM under CentOS 7.

5.2 Main Result

We evaluate the hardware implementation code generated by various baselines across four LLM backbones. Detailed results in Table 1 demonstrate that our proposed ArchCraft achieves superior performance across all metrics, significantly outperforming both direct generation and established agent frameworks. Specifically, while direct LLM generation—serving as a baseline stripped of our neural-symbolic stages—struggles with project-level complexity (achieving only $\sim$30% structural scores), ArchCraft achieves a Paper-level OA of 86.18 and a Code-level OA of 81.48 using the o3-mini backbone. This represents a substantial performance gain of 24.8% and 117.3% over the strongest baselines in architectural fidelity and implementation completeness, respectively. These results validate ArchCraft's robust capability in grounding unstructured academic papers into synthesizable RTL code in a zero-shot manner.

Our analysis indicates that traditional agent frameworks, while outperforming black-box LLMs, often fail at the project level due to their lack of domain-specific architectural grounding. Methods like ChatDev or VerilogCoder generally succeed in reproducing isolated code snippets but diverge significantly from the reference architectures in finer details and interconnect logic. In contrast, our neural-symbolic framework addresses this by formalizing high-level intent into hierarchical graphs before RTL generation. This approach prevents the "logical fragmentation" common in existing methods, where LLMs act as unstructured generators without structural planning. Furthermore, by integrating compiler-level feedback, ArchCraft effectively bridges the gap between abstract specifi-

Method	LLM	Paper-level				Code-level		
		Func.	Struc.	Synth.	**OA**	RTL	TB	**OA**
Direct	Gemini 2.0 Flash	15.00	28.00	27.50	21.45	36.83	25.11	23.78
Direct	Qwen3-Coder-480B	27.50	28.13	28.75	27.88	46.46	43.82	36.11
Direct	GPT-4o	27.22	34.72	34.44	30.94	41.11	31.20	28.92
Direct	o3-mini	31.00	33.50	35.50	32.45	57.33	43.31	40.26
ChatDev	GPT-4o	41.02	44.96	40.10	42.50	17.50	9.92	10.67
VerilogCoder	o3-mini	41.79	38.21	46.07	40.79	43.57	–	17.43
PaperCoder	o3-mini	70.11	69.26	63.07	69.07	50.41	33.93	37.49
ArchCraft	Gemini 2.0 Flash	78.30	77.11	79.67	78.76	57.56	52.06	56.79
ArchCraft	Qwen3-Coder-480B	77.28	80.28	70.13	78.57	66.71	68.40	69.75
ArchCraft	GPT-4o	76.57	76.96	79.21	76.99	72.64	66.26	71.72
ArchCraft	o3-mini	**85.62**	**86.51**	**86.17**	**86.18**	**86.03**	**71.06**	**81.48**

Table 1: Overall evaluation scores on ArchSynthBench, 64 project-level large circuits, over 650 circuit blocks, for Paper-level and Code-level metrics. "Comp." denotes the Compilation Pass Status, "OA" denotes the Over All score. The best performance is denoted in **bold**.

Method	LLM	Paper-level	Code-level
Direct	Gemini 2.0 Flash	22.17	37.03
Direct	Qwen3-Coder-480B	22.00	44.44
Direct	GPT-4o	26.54	47.73
Direct	o3-mini	29.67	50.74
ChatDev	GPT-4o	26.44	10.89
VerilogCoder	o3-mini	31.27	48.89
PaperCoder	o3-mini	49.68	47.33
ArchCraft	Gemini 2.0 Flash	75.24	70.11
ArchCraft	Qwen3-Coder-480B	76.90	78.54
ArchCraft	GPT-4o	75.59	71.47
ArchCraft	o3-mini	79.13	84.33

Table 2: Human expert evaluation scores from both paper and code levels.

Configuration	**Paper-level OA**	**Code-level OA**	**ER (%)**
Direct Gen (Base)	32.45	40.26	12.5
w/o Graphing Phase	54.12	62.85	45.0
w/o Symbolization	76.38	58.42	38.5
ArchCraft (Full)	**86.18**	**81.48**	**93.24**

Table 3: Ablation study on ArchSynthBench. "ER" denotes the Execution Rate.

cations and functionally verifiable hardware. The consistent performance gains across various backbones confirm that the synergy of graphing and symbolization is the key to enhancing automated hardware comprehension from research documentation.

5.3 Ablation Study

To evaluate the individual contributions of the Graphing and Symbolization phases, we conduct an ablation study using o3-mini as the base LLM. We compare the full ArchCraft pipeline against three variants: (1) Direct Gen, bypassing both stages; (2) w/o Graphing, where symbols are extracted directly from text; and (3) w/o Symbolization, where RTL is generated directly from the architectural graph.

As shown in Table 3, the Graphing phase is essential for structural comprehension, as its absence severely degrades Paper-level OA. Meanwhile, the Symbolization phase ensures implementation integrity; without it, even correct architectural understanding fails to translate into executable RTL, as evidenced by the sharp decline in Code-level OA and ER. The full ArchCraft pipeline maximizes both metrics, validating the necessity of this neural-symbolic synergy.

5.4 Human Expert Assessments

To rigorously validate the accuracy of our automated evaluation metrics, we contract three independent human engineering experts to assess the code quality generated by the direct method, other frameworks, and our ArchCraft. To ensure impartiality, we implemented a double-blind protocol where experts were unaware of the specific methodology—ArchCraft or baseline—that produced each code sample. These experts utilized the same evaluation dimensions and scoring rubrics as our machine-based system. As shown in Table 3, the human assessment results indicate that

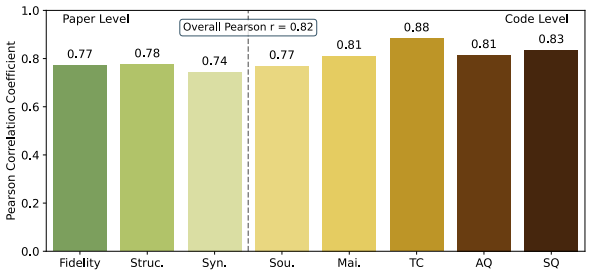

Figure 3: Pearson correlation between human and machine-based scores for ArchCraft.

LLM Model	CD Num.	Mean	Variance
Gemini 2.0 Flash	92	7.58	8.52
Qwen3-Coder-480B	87	6.44	8.47
GPT-4o	70	4.95	9.26
o3-mini	75	2.07	7.64

Table 4: Rectifying phase iteration counts. CD Num. is the abbreviation of Circuit Designs Number.

ArchCraft consistently achieves the highest ranking across all compared methodologies. This expert feedback aligns with the superior performance observed in model-based evaluations, further validating the framework's efficacy. Finally, to confirm the reliability of our automated scoring, we calculated the Pearson correlation between human and machine evaluations. As illustrated in Figure 3, the total Pearson $r = 0.82$, demonstrating strong consistency and justifying the use of our automated evaluation pipeline.

5.5 Hardware Synthesis.

To verify downstream viability, we performed physical synthesis using Synopsys DC on 60 modules sampled via randomized stratified sampling across ArchSynthBench. This evaluation prioritizes the vertical integrity of the ASIC flow; as shown in Table 5, the majority of designs achieved timing closure, confirming the structural integrity of ArchCraft's output. Minor failures, primarily from port mismatches in abstract paper descriptions.

As a case study, Table 6 presents the post-synthesis PPA results made by ArchCraft with o3-mini as backbone, for designs derived from the SPARK (Liu et al., 2024a). All the implementations of the SPARK paper successfully pass synthesis with positive slack, confirming their feasibility for timing closure at the target frequency. Notably, the logic synthesizes a single PE that occupies approximately 4k μm^2, with a power consumption of 0.74 mW. While an area discrepancy exists due to specific design details compared to the original SPARK paper, our generated circuit design reproduces the relative scale and power characteristics expected for compute-intensive accelerators, considering the overall relationships between the various designs. Furthermore, the control and interface logic demonstrates sub-mW power consumption and compact area footprints, consistent with their functional roles within the SPARK architecture. These empirical results underscore the efficacy of our ArchCraft framework in translating high-level architectural descriptions into synthesizable RTL implementations with practical PPA characteristics.

Paper Title	(Success/Total)	Paper Title	(Success/Total)
FlightVGM	6 / 8	PBN	5 / 5
Flex-EGAI	4 / 5	SPARK	7 / 7
LC-MAC	3 / 4	ST-Purning	5 / 5
INSPIRE	4 / 6	MASL_AFU	5 / 7
bitWAVE	5 / 6	ViTA	7 / 8

Table 5: Synthesis success rates across sampled papers.

Design	Power(μW)	Slack/CPD(ns)	Area(μm^2)
control_unit	170.1	3.92/0.93	954.68
im2col_pack_engine	1039.9	2.31/2.52	5758
interface	603.3	3.44/1.41	3945.11
memory_controller	939.7	2.85/2.06	6200.41
pe_array	5262.2	1.04/3.81	32345.91
spark_decoder	109.0	3.88/0.98	568.81
spark_encoder	324.4	3.47/1.38	1676.51

Table 6: Synthesis PPA results for SPARK designs using Synopsys DC. All the designs have no time violations. CPD is the abbreviation of Critical Path Delay.

6 Conclusion

This paper presents a framework, ArchCraft. Unlike traditional methodologies focused on circuit-level optimization, ArchCraft prioritizes the comprehensive understanding of high-level architectural intent by distilling academic prose into hierarchical formal graphs and symbolic specifications. We demonstrate that complex hardware architectures can be captured through a neural-symbolic lens, transforming abstract descriptions into synthesizable projects. To evaluate this capability, we introduced ArchSynthBench, the first benchmark for project-level hardware reproduction from literature. Our results, validated across machine metrics, human expertise, and physical synthesis, confirm that ArchCraft effectively grounds textual concepts into viable logic. We expect the proposed ArchCraft will contribute to accelerating scientific progress by lowering the barriers to hardware replication and enhancing research reproducibility and innovation.

Limitation

(1) Model Dependency: The framework's iterative efficiency is highly sensitive to the choice of the LLM backbone. For example, certain models require significantly more refinement cycles to achieve synthesizable code, leading to varied computational overhead. (2) Physical Discrepancies: Gaps remain between the generated PPA metrics and those reported in original papers. These discrepancies largely stem from the high-level abstraction of academic texts, which often lack the low-level implementation details required for exact physical replication. (3) Synthesis Robustness: While most designs pass verification, occasional synthesis failures still occur due to syntactical errors or port definition mismatches. Improving the framework's ability to autonomously resolve ambiguities in abstract architectural descriptions is necessary for broader hardware design coverage.

Ethic Statement

The primary objective of this work is to automate the replication of experiments detailed in existing, publicly available research papers. While the methodologies themselves are drawn from the public domain, which generally implies transparency, it is important to acknowledge the potential for data leakage associated with the use of our system. Users should therefore be mindful of this possibility, particularly when the replication process might involve sensitive datasets or generate intermediate results that could inadvertently disclose information.

References

Ahmed J Abdelmaksoud, Shady Agwa, and Themis Prodromakis. 2025. Dip: A scalable, energy-efficient systolic array for matrix multiplication acceleration. *IEEE Transactions on Circuits and Systems I: Regular Papers*.

Sami Ben Ali, Silviu-Ioan Filip, and Olivier Sentieys. 2024. A stochastic rounding-enabled low-precision floating-point mac for dnn training. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Andrea Belano, Yvan Tortorella, Angelo Garofalo, Luca Benini, Davide Rossi, and Francesco Conti. 2025. A flexible template for edge generative ai with high-accuracy accelerated softmax & gelu. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*.

Jason Blocklove, Siddharth Garg, Ramesh Karri, and Hammond Pearce. 2023. Chip-chat: Challenges and opportunities in conversational hardware design. In *2023 ACM/IEEE 5th Workshop on Machine Learning for CAD (MLCAD)*, pages 1–6. IEEE.

Oliver Cassidy, Marta Andronic, Samuel Coward, and George A Constantinides. 2025. Reducedlut: Table decomposition with" don't care" conditions. In *Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, pages 36–42.

Chunyun Chen, Lantian Li, and Mohamed M Sabry Aly. 2024a. Vita: a highly efficient dataflow and architecture for vision transformers. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Jiasong Chen, Zeming Xie, Weipeng Liang, Bosheng Liu, Xin Zheng, Jigang Wu, and Xiaoming Xiong. 2024c. Quantization-aware optimization approach for cnns inference on cpus. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 878–883. IEEE.

Kangjoon Choi, Sungho Kim, Jeongmin Kim, and In-Cheol Park. 2024. Hardware-friendly approximation for swish activation and its implementation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 71(10):4516–4520.

Abhay Chopde, Vaibhav Shingde, Tejas Bhagwat, and Sanika Yadav. 2024. Bit-serial risc-v cpu core. In *2024 IEEE Pune Section International Conference (PuneCon)*, pages 1–5. IEEE.

Haotian Cui, Yue Xu, Kuan Pang, Gen Li, Fanglin Gong, Bo Wang, and Bowen Li. 2025. Lumi-lab: a foundation model-driven autonomous platform enabling discovery of new ionizable lipid designs for mrna delivery. *BioRxiv*, pages 2025–02.

Mario Doumet, Marius Stan, Mathew Hall, and Vaughn Betz. 2024. H2pipe: high throughput cnn inference on fpgas with high-bandwidth memory. In *2024 34th International Conference on Field-Programmable Logic and Applications (FPL)*, pages 69–77. IEEE.

Alejandro García-Requejo, Álvaro Hernández, and M Carmen Pérez-Rubio. 2025. Fpga-based soc architecture for real-time device-free localization using a multistatic us sonar. *Measurement*, page 118320.

Xinkuang Geng, Siting Liu, Jianfei Jiang, Kai Jiang, and Honglan Jiang. 2024. Compact powers-of-two: An efficient non-uniform quantization for deep neural networks. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE.

Alireza Ghafarollahi and Markus J Buehler. 2025. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Google DeepMind. 2025. Gemini - our most intelligent ai models. https://deepmind.google/models/gemini/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yufeng Gu, Alireza Khadem, Sumanth Umesh, Ning Liang, Xavier Servot, Onur Mutlu, Ravi Iyer, and Reetuparna Das. 2025. Pim is all you need: A cxl-enabled gpu-free system for large language model inference. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 862–881.

Chia-Tung Ho, Haoxing Ren, and Brucek Khailany. 2025. Verilogcoder: Autonomous verilog coding agents with graph-based planning and abstract syntax tree (ast)-based waveform tracing tool. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 300–307.

Kai Huang, Bowen Li, Siang Chen, Luc Claesen, Wei Xi, Junjian Chen, Xiaowen Jiang, Zhili Liu, Dongliang Xiong, and Xiaolang Yan. 2022. Structured term pruning for computational efficient neural networks inference. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 42(1):190–203.

Longwei Huang, Chao Fang, Qiong Li, Jun Lin, and Zhongfeng Wang. 2024. A precision-scalable risc-v dnn processor with on-device learning capability at the extreme edge. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 927–932. IEEE.

Zhirui Huang, Rui Ma, Shijie Cao, Ran Shu, Ian Wang, Ting Cao, Chixiao Chen, and Yongqiang Xiong. 2025. Tenet: An efficient sparsity-aware lut-centric architecture for ternary llm inference on edge. *arXiv preprint arXiv:2509.13765*.

Dongseok Im and Hoi-Jun Yoo. 2024. Lutein: Dense-sparse bit-slice architecture with radix-4 lut-based slice-tensor processing units. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 747–759. IEEE.

Sunita Jain, Nagaradhesh Yeleswarapu, Hasan Al Maruf, and Rita Gupta. 2024. Memory sharing with cxl: Hardware and software design approaches. *arXiv preprint arXiv:2404.03245*.

Zexi Ji, Hanrui Wang, Miaorong Wang, Win-San Khwa, Meng-Fan Chang, Song Han, and Anantha P Chandrakasan. 2023. A fully-integrated energy-scalable transformer accelerator supporting adaptive model configuration and word elimination for language understanding on edge devices. In *2023 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6. IEEE.

Yiyue Jiang, Andrius Vaicaitis, John Dooley, and Miriam Leeser. 2024. Efficient neural networks on the edge with fpgas by optimizing an adaptive activation function. *Sensors*, 24(6):1829.

Leonardo R Juracy, Alexandre M Amory, and Fernando G Moraes. 2022. A comprehensive evaluation of convolutional hardware accelerators. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(3):1149–1153.

Dongyun Kam, Myeongji Yun, Sunwoo Yoo, Seungwoo Hong, Zhengya Zhang, and Youngjoo Lee. 2025. Panacea: Novel dnn accelerator using accuracy-preserving asymmetric quantization and energy-saving bit-slice sparsity. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 701–715. IEEE.

Alireza Khataei and Kia Bazargan. 2025. Treelut: An efficient alternative to deep neural networks for inference acceleration using gradient boosted decision trees. In *Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, pages 14–24.

Daehan Kwon, Seongju Lee, Kyuyoung Kim, Sanghoon Oh, Joonhong Park, Gi-Moon Hong, Dongyoon Ka, Kyudong Hwang, Jeongje Park, Kyeongpil Kang, and 1 others. 2022. A 1ynm 1.25 v 8gb 16gb/s/pin gddr6-based accelerator-in-memory supporting 1tflops mac operation and various activation functions for deep learning application. *IEEE Journal of Solid-State Circuits*, 58(1):291–302.

V Lalu and 1 others. 2024. Design and implementation of 5-stage pipelined risc-v processor on fpga. In *2024 28th International Symposium on VLSI Design and Test (VDAT)*, pages 1–6. IEEE.

Eunji Lee, Yoonsang Han, and Gordon Euhyun Moon. 2024. Accelerated block-sparsity-aware matrix reordering for leveraging tensor cores in sparse matrix-multivector multiplication. In *European Conference on Parallel Processing*, pages 3–16. Springer.

Lei Lei and Zhiming Chen. 2024. A reconfigurable fused multiply-accumulate for miscellaneous operators in deep neural network. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Richie Li and Sicheng Chen. 2025. Design and implementation of an fpga-based tiled matrix multiplication accelerator for transformer self-attention on the xilinx kv260 som. *arXiv e-prints*, pages arXiv–2503.

Wenjie Li, Aokun Hu, Gang Wang, Ningyi Xu, and Guanghui He. 2022. Low-complexity precision-scalable multiply-accumulate unit architectures for deep neural network accelerators. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(4):1610–1614.

Bo Liu, Anfeng Xue, Ziyu Wang, Na Xie, Xuetao Wang, Zhen Wang, and Hao Cai. 2022a. A reconfigurable approximate computing architecture with dual-vdd for low-power binarized weight network deployment. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(1):291–295.

Fangxin Liu, Ning Yang, Haomin Li, Zongwu Wang, Zhuoran Song, Songwen Pei, and Li Jiang. 2024a. Spark: Scalable and precision-aware acceleration of neural networks via efficient encoding. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 1029–1042. IEEE.

Fangxin Liu, Ning Yang, Zhiyan Song, Zongwu Wang, Haomin Li, Shiyuan Huang, Zhuoran Song, Songwen Pei, and Li Jiang. 2024b. Inspire: Accelerating deep neural networks via hardware-friendly index-pair encoding. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pages 1–6.

Jun Liu, Shulin Zeng, Li Ding, Widyadewi Soedarmadji, Hao Zhou, Zehao Wang, Jinhao Li, Jintao Li, Yadong Dai, Kairui Wen, and 1 others. 2025a. Flightvgm: Efficient video generation model inference with online sparsification and hybrid precision on fpgas. In *Proceedings of the 2025 ACM/SIGDA International Symposium on Field Programmable Gate Arrays*, pages 2–13.

Mingjie Liu, Nathaniel Pinckney, Brucek Khailany, and Haoxing Ren. 2023. Verilogeval: Evaluating large language models for verilog code generation. In *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*, pages 1–8. IEEE.

Shang Liu, Yao Lu, Wenji Fang, Jing Wang, and Zhiyao Xie. 2025b. Sync-llm: Generation of large-scale synthetic circuit code with hierarchical language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17361–17376.

Ye Liu, Fei Wu, Neng Zhao, Qirong Zhang, Wenqiang Wang, Yutong Yang, Xiangting Li, Sixu Li, Zili Huang, Shuang Hao, and 1 others. 2022b. Nvp: A flexible and efficient processor architecture for accelerating diverse computer vision tasks including dnn. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(1):271–275.

Mukul Lokhande, Gopal Raut, and Santosh Kumar Vishvakarma. 2025. Flex-pe: Flexible and simd multi-precision processing element for ai workloads. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Binglei Lou, Richard Rademacher, David Boland, and Philip HW Leong. 2024. Polylut-add: Fpga-based lut inference with wide inputs. In *2024 34th International Conference on Field-Programmable Logic and Applications (FPL)*, pages 149–155. IEEE.

Yingchang Mao, Mingyu Shu, and Qiang Liu. 2024. Pbn: Progressive batch normalization for dnn training on edge device. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Satvik Maurya and Swamit Tannu. 2022. Compaqt: Compressed waveform memory architecture for scalable qubit control. In *2022 55th IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 1059–1077. IEEE.

Zhaoteng Meng, Lin Shu, Jianing Zeng, Zhan Li, Kailin Lv, Haoyue Yang, and Jie Hao. 2024. Masl-afu: A high memory access efficiency 2-d scalable lut-based activation function unit for on-device dnn training. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Zhiwen Mo, Lei Wang, Jianyu Wei, Zhichen Zeng, Shijie Cao, Lingxiao Ma, Naifeng Jing, Ting Cao, Jilong Xue, Fan Yang, and 1 others. 2025. Lut tensor core: A software-hardware co-design for lut-based low-bit llm inference. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture*, pages 514–528.

Gabriele Montanaro, Andrea Galimberti, and Davide Zoni. 2024. A prototype-based framework to design scalable heterogeneous socs with fine-grained dfs. In *2024 IEEE 42nd International Conference on Computer Design (ICCD)*, pages 681–684. IEEE.

Bardia Nadimi, Ghali Omar Boutaib, and Hao Zheng. 2024. Pyranet: A multi-layered hierarchical dataset for verilog. *arXiv preprint arXiv:2412.06947*.

Bardia Nadimi, Ghali Omar Boutaib, and Hao Zheng. 2025. Verimind: Agentic llm for automated verilog generation with a novel evaluation metric. *arXiv preprint arXiv:2503.16514*.

Andre Nakkab, Sai Qian Zhang, Ramesh Karri, and Siddharth Garg. 2024. Rome was not built in a single step: Hierarchical prompting for llm-based chip design. In *Proceedings of the 2024 ACM/IEEE International Symposium on Machine Learning for CAD*, pages 1–11.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng

Ma, and 42 others. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing. *Preprint*, arXiv:2509.22186.

OpenAI. 2024. Hello gpt-4.

OpenAI. 2025. Openai o3-mini. https://openai.com/index/openai-o3-mini/. Accessed: 10 February 2025.

Zhewen Pan, Joshua San Miguel, and Di Wu. 2024. Carat: Unlocking value-level parallelism for multiplier-free gemms. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 167–184.

Gunho Park, Hyeokjun Kwon, Jiwoo Kim, Jeongin Bae, Baeseong Park, Dongsoo Lee, and Youngjoo Lee. 2025. Figlut: An energy-efficient accelerator design for fp-int gemm using look-up tables. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 1098–1111. IEEE.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, and 1 others. 2023. Chatdev: Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Ladan Sayadi, Somayeh Timarchi, and Akbar Sheikh-Akbari. 2023. Two efficient approximate unsigned multipliers by developing new configuration for approximate 4: 2 compressors. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(4):1649–1659.

Samuel Schmidgall and Michael Moor. 2025. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Minju Seo, Jinheon Baek, Seongyun Lee, and Sung Ju Hwang. 2025. Paper2code: Automating code generation from scientific papers in machine learning. *arXiv preprint arXiv:2504.17192*.

Man Shi, Vikram Jain, Antony Joseph, Maurice Meijer, and Marian Verhelst. 2024. Bitwave: Exploiting column-based bit-level sparsity for deep learning acceleration. In *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 732–746. IEEE.

H Umut Suluhan, Serhan Gener, Alexander Fusco, Joshua Mack, Ismet Dagli, Mehmet Belviranli, Cagatay Edemen, and Ali Akoglu. 2024. A runtime manager integrated emulation environment for heterogeneous soc design with risc-v cores. In *2024 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 23–30. IEEE.

Jinwei Tang, Jiayin Qin, Kiran Thorat, Chen Zhu-Tian, Yu Cao, Caiwen Ding, and 1 others. 2024. Hivegen–hierarchical llm-based verilog generation for scalable chip design. *arXiv preprint arXiv:2412.05393*.

Ishraq Tashdid, Dewan Saiham, Nafisa Anjum, Tasnuva Farheen, and Sazadur Rahman. 2025. Ecologic: Enabling circular, obfuscated, and adaptive logic via efpga-augmented socs. *arXiv preprint arXiv:2508.04516*.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Kevin Tien, Ken Inoue, Scott Lekuch, David J Frank, Sudipto Chakraborty, Pat Rosno, Thomas Fox, Mark Yeck, Joseph A Glick, Raphael Robertazzi, and 1 others. 2022. A cryo-cmos transmon qubit controller and verification with fpga emulation. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 13–16. IEEE.

Shreyas Kolala Venkataramanaiah, Jian Meng, Han-Sok Suh, Injune Yeo, Jyotishman Saikia, Sai Kiran Cherupally, Yichi Zhang, Zhiru Zhang, and Jae-Sun Seo. 2023. A 28-nm 8-bit floating-point tensor core-based programmable cnn training processor with dynamic structured sparsity. *IEEE Journal of Solid-State Circuits*, 58(7):1885–1897.

Jure Vreča, Ratko Pilipović, and Anton Biasizzo. 2024. Hardware–software co-design of an audio feature extraction pipeline for machine learning applications. *Electronics*, 13(5):875.

Bowei Wang, Qi Xiong, Zeqing Xiang, Lei Wang, and Renzhi Chen. 2025. Rtlsquad: Multi-agent based interpretable rtl design. *arXiv preprint arXiv:2501.05470*.

Chenyi Wen, Haonan Du, Zhengrui Chen, Li Zhang, Qi Sun, and Cheng Zhuo. 2024. Pace: A piece-wise approximate and configurable floating-point divider for energy-efficient computing. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE.

Jun-Shen Wu and Ren-Shuo Liu. 2023. Fm-p2l: An algorithm hardware co-design of fixed-point msbs with power-of-2 lsbs in cnn accelerators. In *2023 IEEE 41st International Conference on Computer Design (ICCD)*, pages 407–414. IEEE.

Bai-Kui Yan and Shanq-Jang Ruan. 2022. Area efficient compression for floating-point feature maps in convolutional neural network accelerators. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(2):746–750.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chenhao Yang, Jinsheng Bian, Erik Börjeson, and Per Larsson-Edefors. 2025b. Fpga-based wordlength optimization for dsp. In *2025 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Yifan Yang, Joel S Emer, and Daniel Sanchez. 2024. Trapezoid: A versatile accelerator for dense and sparse matrix multiplications. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pages 931–945. IEEE.

Mao-Hsu Yen, Yih-Hsia Lin, Tzu-Feng Lin, Yu-Hui Chen, Yuan-Fu Ku, and Chien-Ting Kao. 2025. Chip design of multithreaded and pipelined risc-v microcontroller unit. *Engineering Proceedings*, 89(1):31.

Zhewen Yu and Christos-Savvas Bouganis. 2024. Auto ws: Automate weights streaming in layer-wise pipelined dnn accelerators. In *2024 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1–6. IEEE.

Danqing Zhang, Baoting Li, Hang Wang, Xuchong Zhang, and Hongbin Sun. 2024. An efficient sparse-aware summation optimization strategy for dnn accelerator. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023a. Repocoder: Repository-level code completion through iterative retrieval and generation. *arXiv preprint arXiv:2303.12570*.

Xiaohui Zhang, Zenghui Yu, Shushan Qiao, and Yi Zhan. 2025. Hardware implementation of block floating-point fft based on approximate computation and conflict-free access. *IEEE Access*.

Yuan Zhang, Lele Peng, Lianghua Quan, Yonggang Zhang, Shubin Zheng, and Hui Chen. 2023b. High-precision method and architecture for base-2 softmax function in dnn training. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 70(8):3268–3279.

Xuanle Zhao, Deyang Jiang, Zhixiong Zeng, Lei Chen, Haibo Qiu, Jing Huang, Yufeng Zhong, Liming Zheng, Yilin Cao, and Lin Ma. 2025a. Vincicoder: Unifying multimodal code generation via coarse-to-fine visual reinforcement learning. *arXiv preprint arXiv:2511.00391*.

Xuanle Zhao, Xianzhen Luo, Qi Shi, Chi Chen, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2025b. Chartcoder: Advancing multimodal large language model for chart-to-code generation. *arXiv preprint arXiv:2501.06598*.

Xuanle Zhao, Zilin Sang, Yuxuan Li, Qi Shi, Weilun Zhao, Shuo Wang, Duzhen Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025c. Autoreproduce: Automatic ai experiment reproduction with paper lineage. *arXiv preprint arXiv:2505.20662*.

Yang Zhao, Di Huang, Chongxiao Li, Pengwei Jin, Ziyuan Nan, Tianyun Ma, Lei Qi, Yansong Pan, Zhenxing Zhang, Rui Zhang, and 1 others. 2024. Codev: Empowering llms for verilog generation through multi-level summarization. *arXiv preprint arXiv:2407.10424*.

Zhongyu Zhao, Rujian Cao, Ka-Fai Un, Wei-Han Yu, Pui-In Mak, and Rui P Martins. 2022. An fpga-based transformer accelerator using output block stationary dataflow for object recognition applications. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 70(1):281–285.

Mengying Zheng, Xianting Yang, Jiaxuan Tang, Yiting Zhu, Xinhui Liu, Jia Zhang, Zhengyuan Chen, Kai Liu, Yun Chen, and Yibo Jiang. 2024. Mg-verilog: Multi-grained dataset towards enhanced llm-assisted verilog generation. *arXiv preprint arXiv:2404.09340*.

Davide Zoni, Andrea Galimberti, and Davide Galli. 2025. An fpga-based open-source hardware-software framework for side-channel security research. *IEEE Transactions on Computers*.

A Ethics Statement

The primary objective of this work is to automate the replication of experiments detailed in existing, publicly available research papers. While the methodologies themselves are drawn from the public domain, which generally implies transparency, it is important to acknowledge the potential for data leakage associated with the use of our system. Users should therefore be mindful of this possibility, particularly when the replication process might involve sensitive datasets or generate intermediate results that could inadvertently disclose information.

B Prompts

In this section, we provide a detailed exposition of the four phases within our framework. The specific prompts utilized at each stage of this process are presented in Section B.1 is Graphing Prompts, Section B.2 is Symbolize Prompts, Section B.3 is Coding Prompts, including original coding prompts and rectifying.

B.1 Graphing Phase

We show 4 steps' prompts in our Graphing Phase. Figure 6 is Define Graph Scope and Theme. Figure 7 is Define Graph Nodes and Static Links. Figure 8 is Plan Graph Direction and Internal Node Properties. Figure 9 is Add Global Constraints.

B.2 Symbolize Phase

We show 3 steps' prompts in our Graphing Phase. Figure 10 is Interface Protocol Design, Internal Logic Design,and Verification & Implementation Plan.

B.3 Coding Phase

In this section, we demonstrated the decoupling generation of RTL and TB code in Figure 11, and the prompt used in the rectifying process of the generated code in Figure 12.

C Input Modalities

To verify the impact of visual information on hardware reproduction, we conducted an ablation study on a representative subset of 10 papers from ArchSynthBench. We compared our standard Text-Only pipeline (using MinerU for conversion) against a Multi-modal pipeline where both the processed text and architectural diagrams (processed via a Vision-Language Model) were provided to the Graphing Agent. The results in Table 7 show that the Paper-level Score —which measures architectural fidelity—improved by only 0.5% with the addition of visual inputs. This confirms that for high-quality ASIC papers, the rigorous natural language descriptions of logic and functional specifications are the primary drivers for successful reproduction.

Input Modality	Paper-level	Code-level	Synthesis
Text-Only	85.8	81.2	92.0%
Multi-modal	86.3	81.5	92.5%
Difference	+0.5	+0.3	+0.5%

Table 7: Comparison of Text-Only vs. Multi-modal inputs across 10 project-level circuits. The marginal improvement suggests that textual logic analysis in hardware papers is sufficient for high-fidelity reconstruction.

D Showcase of Symbolic

To further clarify the mechanics of ArchCraft, Figure 4 illustrates the transformation pipeline using the SPARK decoder as a representative module. The process begins with the Graphing Agent, which distills unstructured textual descriptions—such as logic for high and low-precision identifiers—into a structured Architectural Blueprint. This blueprint identifies core hardware nodes, their hierarchy, and static data-flow links. Building on this structural foundation, the Symbolize Agent translates these nodes and attributes into a rigorous SIR. This IR formalizes natural language logic into precise transition and assignment rules, such as $(ident = 0) \rightarrow (data_out' = data_in[3:0] \wedge en_out' = 0)$, effectively bridging the gap between high-level intent and hardware-aware constraints. Finally, the Coding Agent maps this symbolic logic directly into synthesizable Verilog syntax. By ensuring the RTL implementation is strictly grounded in the formal symbolic specification, this methodology guarantees that the generated circuit maintains high structural integrity and meets the timing and power characteristics intended in the original architecture.

E Differences between Paper and Document

In this subsection, we provide a detailed comparison of the differences 8 between academic papers and industrial design documents as inputs to the ArchCraft framework. This distinction repre-

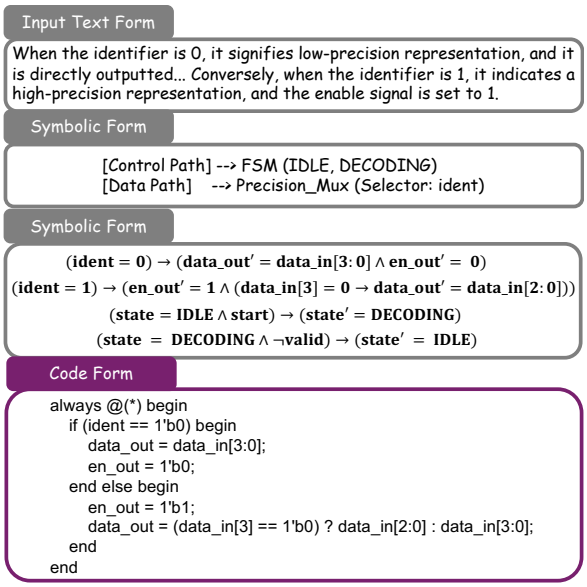

Figure 4: Detailed showcase of the transition from the abstract textual description of the SPARK paper to its SIR and the final synthesizable Verilog code.

sents the foundational starting point for the entire ArchCraft process and the design motivation that led to the creation of its neural-symbolic Graph-based Framework.

F Scoring Criteria

To support the quantitative and qualitative analyses presented in the main text, this appendix provides the detailed evaluation rubrics used to assess the quality of the generated hardware designs. These rubrics are explicitly divided into two parts: criteria for the design implementation (RTL) and criteria for its verification environment (Testbench).

Table 9 details the evaluation rubric for the **RTL Design**. This rubric is structured around five core dimensions:

- **Functional Correctness:** Assesses whether the design accurately implements the required algorithms and logic.
- **Robustness:** Examines the design's handling of boundary conditions, exceptions, and timing issues (e.g., CDC).
- **Structural Fidelity:** Evaluates the code's modularity, structural clarity, and naming conventions.
- **Synthesis Compatibility:** Ensures the code is synthesizable and evaluates the reasonableness of its resource utilization.
- **Readability:** Assesses the quality of comments and documentation, which relates to design maintainability.

Correspondingly, **Table 10** presents the evaluation rubric for the **TB Design**. This rubric focuses on four key aspects:

- **Test Case Coverage:** Evaluates whether the stimulus adequately covers functional paths, boundaries, and corner cases.
- **Robustness:** Focuses on whether the assertions are accurate, effective, and stable under various stimuli.
- **Stimulus Generation Quality:** Assesses the use of directed, random, and constrained-random strategies.
- **Maintainability:** Examines the clarity, extensibility, and debuggability of the testbench code itself.

As mentioned in the main text, each dimension is scored out of 100 points. The specific questions listed within these tables provide the basis for fine-grained scoring and qualitative feedback by human evaluators.

G Paper in ArchSynthBench Details

In this subsection, we detail the corpus of academic papers that constitutes the **ArchSynthBench** benchmark. These papers serve as the high-level, abstract design specifications that our framework, ArchCraft, aims to interpret and implement. The complete list of these papers is presented in **Table 12**, which is organized to highlight the breadth and timeliness of our curated corpus.

The table provides three key pieces of information for each entry:

- **Year:** The publication year of the paper. This chronological organization (spanning 2022 to 2025) demonstrates that the benchmark is contemporary and built upon state-of-the-art research in hardware acceleration.
- **Paper:** The common acronym or a short name for the paper, along with its bibliographic citation key (e.g., `(Zhao et al., 2022)`). This provides full transparency and traceability, allowing researchers to reference the original design specifications used in our evaluation.
- **Theme** A set of descriptive terms summarizing the paper's core technical contribution, target application, or key architectural paradigm.

An analysis of the **Theme** column reveals the extensive diversity of our benchmark. The corpus

Feature / Attribute		Academic Paper	Industrial Doc.
Accessibility	Publicly Searchable	✓	✗
	Abundant in Quantity	✓	✗
	Internally Confidential	✗	✓
Content Focus	Core Algorithm / Theory	✓	(✓)
	Rigorous Validation / Proofs	✓	✗
	Detailed Circuit-level Specs	✗	✓
	Implementation Details (APIs, etc.)	✗	✓
	Edge Cases	✗	✓
Form	Standardized Structure	✓	✗
	High-Level Abstraction	✓	✗
	Low-Level Granularity	✗	✓
Lifecycle	Static / Archived	✓	✗
	Dynamic / Living Document	✗	✓

Table 8: Feature-based Comparison of Academic Papers and Industrial Design Documents.

Functional Correctness	• Does the RTL correctly implement the key algorithms, logic, or functional points described in the logical analysis? Is the functionality complete?
	• Does it faithfully reproduce the architecture, dataflow, and bit-width details specified in the logical analysis?
Robustness	• Does the RTL handle all input cases, boundary conditions, and potential exceptions (e.g., divide-by-zero, overflow, illegal inputs)?
	• Is the reset logic clear and effective? Is Clock Domain Crossing (CDC) handled correctly (if applicable)?
	• Is the design complete, unambiguous, and synthesizable without errors?
Structural Fidelity	• Is the code structure clear? Does it adhere to the logical layering from the analysis (if mentioned)?
	• Is the module partitioning reasonable and the naming convention consistent?
Synthesis Compatibility	• Can the design be successfully synthesized by mainstream tools? Does it avoid non-synthesizable constructs?
	• Is the resource utilization reasonable and aligned with the goals set in the logical analysis?
Readability	• Does the code have sufficient comments? Are signal names clear and meaningful?
	• Does it include auxiliary documentation (e.g., parameter descriptions, structural comments)?

Table 9: Evaluation Rubric for RTL Design (100 points per dimension)

Test Case Coverage	• Does the TB generate sufficient and representative test cases to cover all functional paths of the RTL?
	• Does it cover boundary conditions, typical values, random values, and corner cases of the input space?
	• Does it consider various timing combinations of control signals (e.g., reset, enable)?
Robustness	• Do the assertions (‘assert‘) accurately reflect the RTL’s expected behavior?
	• Are assertions granular enough to effectively capture errors in outputs or internal states?
	• Are assertions robust, remaining stable and correct under varying or randomized stimulus?
	• Are there checks on key intermediate signals, not just the final outputs?
Stimulus Generation Quality	• Is the method for generating input stimulus logical and flexible?
	• Does it use a mix of directed, random, and constrained-random stimulus to enhance coverage?
	• Is the generation of clock and reset signals compliant with the design specificaåtion?
Maintainability	• Is the testbench code clear, well-structured, and easy to understand?
	• Is it easy to modify, extend with new test cases, or debug?
	• Is the code free of redundancy and unnecessary complexity?

Table 10: Evaluation Rubric for TB Design (100 points per dimension)

was intentionally curated to move beyond simple, canonical designs and to reflect the complexity of the modern hardware landscape. The topics span:

- **Classic Acceleration Domains:** Such as CNN Accelerators (`CNN Accelerator`), Approximate Computing (`Approximate Computing / BNN`), and efficient inference (`Efficient NN Inference`).
- **Modern Neural Architectures:** Including dedicated hardware for Transformers (`FPGA Transformer`), Vision Transformers (`Vision Transformer / Accelerator`), and Large Language Models (`LLM Inference / PIM / CXL`).
- **Specific Hardware Techniques:** Such as `LUT-based NN`, `Sparsity`, `Quantization`, and Accelerator-in-Memory (`AiM`).
- **Emerging Paradigms:** The benchmark even includes forward-looking topics like CXL-based memory systems (`Memory Systems / CXL`) and Quantum Computing control (`Quantum Computing / Control`).

In summary, the paper corpus detailed in Table 12 provides a rich, challenging, and representative set of tasks. This diversity is essential for robustly evaluating the capability of any high-level synthesis agent to generalize across different domains and effectively bridge the gap from abstract textual specifications to concrete hardware implementations.

H Code-Level Evaluation Scores

In Table 1, five key metrics are folded into RTL and TB, and we record detailed scores in all dimensions for code-level in Table 11.

I Original Scores

We provide the score lists from both machine and human evaluations as illustrative examples. Given that an exhaustive presentation of all circuit modules is infeasible, we selectively present the scores

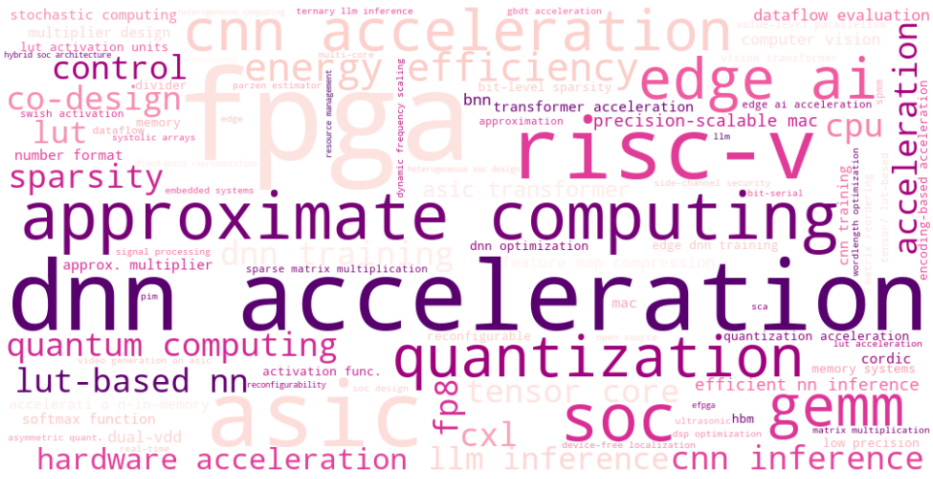

Figure 5: Theme of the papers.

for nearly 100 circuit modules derived from the 10 papers in batch1 of ArchSynthBench, as evaluated by othermethod-o3-mini and archcraft-o3-mini.

Specifically, we provide the LLM evaluation results of Direct-o3-mini from Table 1, presented in Table 13, and the human evaluation results from Table 3, presented in Table 14; we also provide the LLM evaluation results of ArchCraft-o3-mini from Table 1, presented in Tables 15 and 16, as well as the human evaluation results from Table 3, presented in Table 17.

J Analysis

Rectifying Analysis. Our analysis focuses on the mean and variance of feedback-based rectification steps for code errors within the ArchCraft framework, utilizing Gemini 2.0 Flash, Qwen3-Coder-480B, GPT-4o (OpenAI, 2024), and o3-mini. In Table 4, Gemini 2.0 Flash required the most iterations (average 7.58). Conversely, o3-mini had the lowest average and minimal variance, suggesting more stable outputs. Our framework enhances all LLMs and that o3-mini's generated RTL has superior initial correctness and lower integration costs.

K Physical Implementation

In this section, we show the DC results of the final case study in the main text. Three designs are shown. Figures 13, Figure 14, Figure 15 show the PPA of the pe_array, Figure 16, Figure 17, Figure 18 show the PPA of the spark_encoder, Figure 19, Figure 20, Figure 21 show the PPA of the spark_decoder.

L Computational Cost and Efficiency Analysis

To provide a transparent assessment of the operational overhead, we analyze the computational resources and token consumption required by ArchCraft. All statistics are averaged across the 64 project-level circuits in ArchSynthBench, utilizing $o3-mini$ as the primary backbone.

The end-to-end transformation of a typical 10-page academic PDF into verified RTL implementation entails a total consumption of approximately 90k input tokens and 30k output tokens. This total includes the initial architectural distillation in the Graphing Phase, the formal symbolic definition in the Symbolize Phase, and the subsequent iterative refinement within the Coding and Compiling loops.

Method	LLM	RTL		TB		
		Robust.	Read.	Cov.	Assert.	Stim.
Direct	Gemini 2.0 Flash	30.33	43.33	23.61	19.11	32.61
Direct	Qwen3-Coder-480B	42.08	50.83	46.74	41.74	42.99
Direct	GPT-4o	32.50	49.72	35.27	23.61	34.72
Direct	o3-mini	52.33	62.33	53.71	31.11	45.11
ChatDev	GPT-4o	17.08	17.92	17.63	1.11	11.03
VerilogCoder	o3-mini	34.62	52.50	00.00	00.00	00.00
PaperCoder	o3-mini	51.07	49.76	45.20	20.81	35.78
ArchCraft	Gemini 2.0 Flash	53.38	61.73	61.16	36.73	58.30
ArchCraft	Qwen3-Coder-480B	62.91	70.51	72.25	**64.44**	68.51
ArchCraft	GPT-4o	68.25	77.03	69.50	60.93	68.37
ArchCraft	o3-mini	**80.95**	**91.13**	**75.81**	62.58	**74.77**

Table 11: The code-level evaluation is based on five key metrics. The best performance is denoted in **bold**.

Specifically, the majority of the token budget is allocated to the iterative rectifying phase, as the framework autonomously invokes the LLM to resolve syntax or logic discrepancies reported by the EDA toolchain.

In terms of temporal efficiency, the average wall-clock time for generating a complete, synthesizable project is approximately 240 minutes. This duration encompasses the neural inference time for all agentic interactions and the execution latency of the symbolic verification and Verilator simulation cycles. Despite these computational requirements, ArchCraft represents a paradigm shift in productivity: the entire automated process is completed within minutes, whereas manual reproduction of similar closed-source hardware architectures typically demands 2–4 weeks of expert engineering effort. Consequently, the marginal cost of LLM tokens is negligible compared to the significant reduction in human labor and the acceleration of the research design cycle.

Year	Paper Name	
2022	OBS-TA (Zhao et al., 2022)	Efficient Compression (Yan and Ruan, 2022)
	LC-MAC (Li et al., 2022)	Comprehensive Evaluation (Juracy et al., 2022)
	ST-Purning (Huang et al., 2022)	Approx. Arch. (Liu et al., 2022a)
	COMPAQT (Maurya and Tannu, 2022)	NVP (Liu et al., 2022b)
	Cryo-CMOS Transmon (Tien et al., 2022)	GDDR6-Based AiM (Kwon et al., 2022)
2023	AFIES (Ji et al., 2023)	Accurate Binary-Stochastic (Zhang et al., 2023b)
	Efficient Multipliers (Sayadi et al., 2023)	High-Precision Softmax (Zhang et al., 2023b)
	FP Tensor Core (Venkataramanaiah et al., 2023)	FM-P2L (Wu and Liu, 2023)
2024	INSPIRE (Liu et al., 2024b)	PACE (Wen et al., 2024)
	PBN (Mao et al., 2024)	PolyLUT-Add (Lou et al., 2024)
	SPARK (Liu et al., 2024a)	RFMA (Lei and Chen, 2024)
	BitWave (Shi et al., 2024)	Precision-Scalable (Huang et al., 2024)
	MASL-AFU (Meng et al., 2024)	Trapezoid (Yang et al., 2024)
	AESA (Zhang et al., 2024)	ViTA (Chen et al., 2024a)
	AutoWS (Yu and Bouganis, 2024)	Carat (Pan et al., 2024)
	Memory Sharing with CXL (Jain et al., 2024)	Block-Sparsity (Lee et al., 2024)
	ENN (Jiang et al., 2024)	Quantization-aware (Chen et al., 2024c)
	CPoT (Geng et al., 2024)	LUTein (Im and Yoo, 2024)
	FPMAC (Ali et al., 2024)	TreeLUT (Khataei and Bazargan, 2025)
	H2PIPE (Doumet et al., 2024)	Approx. for Swish (Choi et al., 2024)
	Fine-grained Dfs (Montanaro et al., 2024)	Bit-Serial (Chopde et al., 2024)
	Heterogeneous SoC (Suluhan et al., 2024)	5-Stage (Lalu et al., 2024)
	Co-Design (Vreča et al., 2024)	
2025	Flex-EGAI (Belano et al., 2025)	FIGLUT (Park et al., 2025)
	FlightVGM (Liu et al., 2025a)	TENET (Huang et al., 2025)
	PIM Is All You Need (Gu et al., 2025)	Panacea (Kam et al., 2025)
	RLUT (Cassidy et al., 2025)	Flex-PE (Lokhande et al., 2025)
	LUTTC (Mo et al., 2025)	FPGA-based SoC (García-Requejo et al., 2025)
	Side-Channel Security (Zoni et al., 2025)	DiP (Abdelmaksoud et al., 2025)
	Wordlength (Yang et al., 2025b)	ECOLogic (Tashdid et al., 2025)
	SoM (Li and Chen, 2025)	Microcontroller (Yen et al., 2025)
	Floating-Point FFT (Zhang et al., 2025)	

Table 12: List of Papers

Design Name	Paper-level			Code-level				
	Fidelity	Structural	Synth.	RTL		TB		
				Sou.	Mai.	TC	AQ	SQ
Flex-EGAI	4.0	2.0	5.0	5.0	6.0	5.0	3.0	4.0
OBS-TA	4.0	4.0	3.5	6.0	6.5	4.6	5.0	4.5
INSPIRE	4.0	4.0	4.0	5.5	6.5	5.5	2.0	4.5
LC-MAC	4.0	4.5	4.0	5.5	6.5	5.0	2.0	4.0
PBN	3.0	3.5	2.5	4.5	6.0	4.5	0.0	2.0
SPARK	2.0	3.0	4.0	5.0	6.0	5.0	3.0	5.0
BitWave	3.0	4.0	3.5	4.5	6.0	5.0	3.0	4.5
FlightVGM	1.0	2.0	3.0	4.5	6.0	6.0	5.5	5.0
MASL-AFU	3.0	3.0	3.0	5.5	6.0	6.0	2.0	4.5
ST-Purning	3.0	3.5	3.0	5.5	6.0	6.0	4.5	6.0

Table 13: Machine Scoring Results Using o3-mini as the LLM in the Direct Method.

Paper Name	Paper-level			Code-level				
	Fidelity	Structural	Synth.	RTL		TB		
				Sou.	Mai.	TC	AQ	SQ
INSPIRE	2.0	3.0	9.0	5.0	6.0	5.0	2.0	4.0
MASL-AFU	3.0	2.0	8.0	5.0	6.0	6.0	2.0	4.0
SPARK	2.0	2.0	9.0	5.0	5.0	4.0	3.0	5.0

Table 14: Human Expert Scoring Results Using o3-mini as the LLM in the Direct Method.

Paper Name	Design Name	Paper-level			Code-level				
		Fidelity	Structural	Synth.	RTL		TB		
					Sou.	Mai.	TC	AQ	SQ
Flex-EGAI	softex_streamer	7.0	8.0	8.0	9.0	8.0	8.0	5.0	8.0
	top	5.0	6.0	7.0	5.0	8.0	5.0	2.0	5.0
	tpu	9.0	8.0	8.0	8.0	9.0	7.0	6.0	7.0
	riscv_cluster	9.0	9.0	9.0	8.0	9.0	7.0	5.0	7.0
	softex_datapath	9.0	9.0	8.0	5.0	6.0	6.0	5.0	5.0
	softex_control	9.0	8.0	9.0	8.0	9.0	8.0	6.0	8.0
	memory_controller	8.0	8.0	8.0	8.0	9.0	7.0	5.0	7.0
OBS-TA	config	10.0	9.0	9.0	10.0	10.0	8.0	8.0	8.0
	obs_datapath	8.0	8.5	8.0	7.5	9.0	7.5	5.0	7.5
	top	6.0	7.5	8.0	7.0	7.5	8.0	7.0	7.5
	obs_controller	9.0	10.0	9.5	8.5	10.0	7.5	6.5	7.5
	nonlinear_block	6.0	6.0	7.5	5.0	7.5	6.0	5.0	6.0
	mem_controller	9.0	8.5	9.0	4.5	7.0	6.0	4.5	6.0
INSPIRE	accumulation_unit	9.0	9.0	9.0	8.5	9.5	8.0	5.0	8.0
	activation_encoder	9.0	9.0	9.0	8.5	9.0	9.0	8.0	7.5
	control_unit	9.0	9.5	9.0	5.5	6.5	4.5	3.0	5.5
	ip_pe	9.0	9.5	9.0	9.0	9.5	10.0	9.0	9.5
	ip_pe_array	9.5	9.5	10.0	9.0	9.5	8.0	7.0	7.5
	memory_interface	9.0	9.0	9.0	8.0	9.0	8.0	7.0	8.0
	top	8.0	9.0	8.5	7.5	9.0	7.0	5.0	7.5
	weight_buffer	9.0	9.0	10.0	5.5	6.0	4.5	5.0	4.5
LC-MAC	bit_brick	8.0	9.0	8.5	9.0	9.0	9.0	8.5	8.0
	rfu_core	7.5	9.0	8.5	8.0	8.0	8.5	6.0	7.5
	shift_add_unit	9.0	9.0	9.0	9.0	9.5	7.5	8.0	7.5
	bbu_unit	9.5	9.5	9.0	9.0	9.5	7.5	5.0	7.5
	top	9.0	9.0	9.0	8.0	9.5	9.0	6.0	8.5
	mode_controller	9.0	9.5	10.0	9.5	9.5	8.0	7.5	8.0
PBN	pbn_control_unit	9.0	9.0	9.0	8.5	9.0	8.5	5.0	7.5
	pbn_datapath	8.0	7.0	7.5	7.5	9.0	8.0	5.0	7.5
	pbn_fifo	10.0	9.0	10.0	9.0	9.5	8.5	8.0	9.0
	pbn_nps	8.0	9.0	8.0	7.5	9.0	7.0	5.0	7.0
	pbn_scs	9.0	9.0	8.5	8.0	9.0	7.0	5.0	7.0
	pbn_top	9.0	10.0	9.0	8.5	10.0	7.0	6.0	7.5

Table 15: Machine Scoring Results Using o3-mini as the LLM in the ArchCraft. I

Paper Name	Design Name	Paper-level			Code-level				
		Fidelity	Structural	Synth.	RTL		TB		
					Sou.	Mai.	TC	AQ	SQ
SPARK	control_unit	9.0	9.0	9.0	9.0	10.0	8.0	9.0	8.0
	im2col_pack_engine	9.0	9.5	10.0	9.0	9.5	7.5	5.0	8.5
	interface	8.5	8.0	9.0	8.0	8.0	6.0	6.0	7.0
	memory_controller	9.0	9.0	9.5	8.5	9.0	7.0	6.0	7.0
	pe_array	8.0	9.0	9.0	9.0	9.5	7.0	5.0	7.0
	spark_decoder	10.0	10.0	10.0	10.0	10.0	9.0	5.0	7.5
	spark_encoder	9.0	9.0	9.5	9.5	10.0	8.0	7.0	7.0
	top	9.0	9.0	9.0	8.0	9.0	7.5	7.0	7.0
BitWave	bce_array	9.0	9.0	8.5	8.5	9.5	9.0	8.0	9.0
	control_unit	9.0	8.5	9.0	8.0	9.0	8.0	7.5	8.0
	data_dispatcher	9.0	8.0	9.0	8.0	9.0	7.0	6.0	7.0
	memory_interface	8.0	9.0	9.0	7.5	9.0	7.0	6.0	7.0
	output_formatter	10.0	10.0	9.5	9.5	9.0	9.0	7.5	8.5
	top	8.0	9.0	9.0	7.5	8.5	8.0	5.0	8.0
	zcip_parser	9.5	9.5	10.0	9.0	9.0	8.0	7.0	8.0
FlightVGM	matrix_processing_engine	6.0	6.0	7.0	5.0	7.0	6.0	4.0	5.0
	cpu_scheduler	9.0	8.5	9.0	8.5	9.0	8.5	7.5	8.0
	top	5.0	4.0	7.5	6.0	7.0	6.0	4.0	5.0
	recovery_unit	9.0	8.5	9.0	9.0	10.0	8.5	8.0	8.5
	global_interconnect	7.0	9.0	9.0	7.5	10.0	9.0	9.0	9.0
	sparsification_unit	9.0	8.5	8.0	8.5	9.0	7.0	5.0	7.0
	dsp_e	8.0	8.5	8.0	7.5	9.0	6.0	6.0	7.0
	mmu_controller	9.0	8.5	9.0	8.5	9.5	8.0	7.5	7.5
	compute_core	6.0	4.0	6.0	7.0	8.0	7.0	5.0	6.0
	special_function_unit	8.0	8.0	7.5	7.5	8.5	7.0	5.0	6.0
MASL-AFU	bsearch_unit	9.0	8.0	9.0	9.0	8.0	8.0	7.0	8.0
	control_unit	9.0	9.0	9.0	8.0	8.5	7.5	7.0	7.5
	lut_unit	9.0	9.0	9.0	8.5	9.0	8.5	8.0	9.0
	mac_unit	9.0	9.0	8.0	8.5	9.5	8.5	8.0	7.5
	masl_afu_top	9.0	10.0	9.0	9.0	9.5	8.0	6.0	8.0
	masl_afu_unit	9.0	9.0	9.0	8.0	9.5	7.5	6.0	7.5
	memory_controller	9.0	8.5	9.0	8.0	9.0	8.0	6.0	8.0
	scalability_node	9.0	8.5	9.0	9.0	9.5	9.0	8.0	9.0
	shared_buffer	9.0	9.0	8.5	8.5	9.5	9.0	7.5	8.0
	top	9.0	9.5	9.0	9.0	10.0	6.5	5.0	6.0
ST-Pruning	accumulator	9.0	9.0	9.0	9.0	10.0	9.0	8.0	8.0
	config_interface	8.0	9.0	9.0	8.5	9.0	7.5	5.0	7.5
	control_unit	9.0	9.0	9.0	8.5	9.5	8.0	5.0	7.5
	fsde_encode	9.0	7.5	9.0	8.0	9.0	9.0	7.0	8.0
	gpe_array	9.0	8.5	8.0	5.5	6.0	4.0	4.0	4.0
	memory_controller	9.0	9.0	9.0	8.5	9.0	7.5	7.5	7.5
	top	8.0	9.0	8.5	7.5	9.0	7.0	6.0	7.5

Table 16: Machine Scoring Results Using o3-mini as the LLM in the ArchCraft. II

Paper Name	Design Name	Paper-level			Code-level				
		Fidelity	Structural	Synth.	RTL		TB		
					Sou.	Mai.	TC	AQ	SQ
INSPIRE	accumulation_unit	10.00	7.33	8.50	9.17	9.00	8.33	5.67	8.33
	activation_encoder	9.17	7.67	8.33	9.00	9.17	9.33	8.67	7.50
	control_unit	5.50	8.33	5.67	5.00	5.50	4.33	3.33	5.17
	ip_pe	9.17	9.67	9.00	9.33	9.17	10.00	9.33	9.67
	ip_pe_array	9.00	10.00	9.50	9.67	9.00	8.67	7.33	8.50
	memory_interface	7.67	8.50	9.17	8.33	8.67	7.50	6.17	7.33
	top	7.33	7.50	7.67	7.17	7.33	5.50	4.67	4.50
	weight_buffer	5.17	5.33	7.50	5.67	6.17	4.67	5.50	4.17
SPARK	control_unit	9.00	9.50	9.17	9.33	10.00	8.67	9.33	8.50
	im2col_pack_engine	10.00	9.33	10.00	9.17	10.00	7.33	5.17	8.67
	interface	8.33	7.50	8.17	8.00	8.33	6.67	6.50	7.17
	memory_controller	9.17	9.00	9.50	9.17	9.67	7.50	5.67	7.33
	pe_array	8.67	9.17	9.00	9.00	9.50	7.17	5.50	7.50
	spark_decoder	10.00	10.00	10.00	10.00	10.00	9.33	5.67	7.67
	spark_encoder	10.00	10.00	10.00	10.00	10.00	8.50	7.17	7.50
	top	8.50	8.17	8.33	8.67	8.00	7.67	7.33	7.67
MASL-AFU	bsearch_unit	9.33	8.50	9.17	9.00	8.67	8.33	7.67	7.50
	control_unit	9.17	9.00	9.50	9.67	9.33	7.50	7.17	7.67
	lut_unit	7.50	7.67	8.33	8.50	8.00	8.17	8.67	8.50
	mac_unit	9.00	9.50	8.17	8.33	10.00	8.67	8.50	8.33
	masl_afu_top	9.67	8.33	8.67	8.50	9.17	8.00	6.33	7.17
	masl_afu_unit	9.50	9.17	8.50	8.00	10.00	7.67	7.50	7.33
	memory_controller	8.67	8.50	9.33	9.50	9.00	8.33	7.17	7.67
	scalability_node	9.00	7.67	10.00	10.00	10.00	9.50	8.33	9.17
	shared_buffer	9.17	8.33	9.00	9.00	9.50	7.17	5.50	8.67
	top	8.50	9.00	8.67	8.17	9.00	7.50	5.67	7.17

Table 17: Overall Human Scoring Results Using o3-mini as the LLM in ArchCraft.

Prompts for graphing papers. I

System Role Prompt:
You are an expert hardware engineer and RTL designer with a deep understanding of digital design and hardware implementation.
You will receive a research paper in <paper_format> format.
Your task is to create a detailed and efficient plan to implement the hardware described in the paper. This plan should align precisely with the paper's hardware architecture, timing requirements, and performance metrics.
The plan must be clear, structured, and focused on hardware implementation details.
Task Prompt:
1. We want to implement the hardware described in the attached paper.
2. The authors did not release any official RTL code, so we have to plan our own implementation.
3. Before writing any Verilog code, please outline a comprehensive plan that covers:
- Key details from the paper's Hardware Architecture.
- Important aspects of Implementation, including module hierarchy, interfaces, timing requirements, and verification strategy.
4. The plan should be as detailed and informative as possible to help us write the final RTL code later.
Instruction Prompt:
The response should give us a strong roadmap for hardware implementation, making it easier to write the RTL code later.

Figure 6: Prompts for graphing papers.

Prompts for graphing papers. II

General Prompt:
Your goal is to create a concise, usable, and complete hardware system design for implementing the paper's method. Use appropriate hardware design practices and keep the overall architecture modular.
Based on the plan for implementing the paper's main method, please design a concise, usable, and complete hardware system. Keep the architecture modular and make effective use of standard hardware design patterns.
Format Prompt:
Please note that this is just an example and in the processing of the actual module, there is no need to strictly adhere to the file names and the number of modules therein. Please refer to the actual analysis results.
Action Prompt:
Follow the instructions of the node to generate the output and ensure that it adheres to the format example. Please remember that the file list and module names should be adjusted according to the actual designed functions and do not have to strictly follow the examples.

Figure 7: Prompts for graphing papers.

Prompts for graphing papers. III

General Prompt:
Your goal is to break down tasks according to PRD/technical design, generate a task list, and analyze task dependencies.
You will break down tasks, analyze dependencies.
You outline a clear PRD/technical design for implementing the paper's hardware method.
Now, let's break down tasks according to PRD/technical design, generate a task list, and analyze task dependencies.
The Logic Analysis should not only consider the dependencies between modules but also provide detailed descriptions to assist in writing the RTL code needed to implement the paper.
Format Prompt:
Please note that this is just an example and in the processing of the actual module, there is no need to strictly adhere to the file names and the number of modules therein. Please refer to the actual analysis results
Action Prompt:
Follow the node instructions above, generate your output accordingly, and ensure it follows the given format example.

Figure 8: Prompts for graphing papers.

Prompts for graphing papers. IV

General Prompt:
You write elegant, modular, and maintainable RTL code. Adhere to hardware design guidelines. Based on the paper, plan, and design specified previously, follow the "Format Example" and generate the code. Extract the hardware details from the above paper (e.g., clock frequency, data width, memory size, etc.), follow the "Format example" and generate the code. DO NOT FABRICATE DETAILS — only use what the paper provides. You must write 'config.yaml'.

Figure 9: Prompts for graphing papers.

Prompts for symbolize. I

System Role Prompt: You are an expert hardware interface designer. Your task is to take a high-level module definition and design its detailed physical interfaces and protocols.
Task Prompt: Based on the high-level plan and the module list, refine the interfaces for the module: [Module_Name_From_Graphing].
You must define:
Handshake Protocols: Specify the exact valid/ready or ack/req logic for each data port.
Memory Interfaces: If this module connects to memory, define the protocol (e.g., AXI-Lite, AXI-Stream, or simple BRAM interface) including all necessary signals (ar_addr, ar_valid, aw_addr, w_data, b_ready, etc.).
Detailed Timing: Specify any multi-cycle signal behaviors or interface timing constraints not covered by the global config.
Control Signals: Define any new control signals required for this specific module's operation (e.g., start_processing, op_mode, done_tick).

Prompts for symbolize. II

System Role Prompt: You are an expert RTL micro-architect. Your task is to design the internal control and data path logic for a given hardware module based on its function and interfaces.
Task Prompt: Design the internal micro-architecture for module:[Module_Name_From_Graphing].
You must provide two separate but related designs:
State Machine (FSM) Design:
Identify FSMs: List all required Finite State Machines.
Define States: For each FSM, list its states.
Define Transitions: Describe the logic that causes transitions between states.
Control Logic: Specify the control signals generated in each state.
Reset Behavior: Define the FSM's state upon rst_n assertion.
Data Path Design:
Data Flow: Describe the path data takes through the module, from input ports to output ports.
Key Components: List the necessary data path components.
Pipeline Stages: If pipelining is required, define the logic for each stage.
Data Formats: Specify any changes in data format.

Prompts for symbolize. III

System Role Prompt: You are an expert verification engineer and physical designer. Your task is to create a test plan and define implementation constraints for a given hardware module.
Task Prompt: Create the verification plan and define implementation constraints for module: [Module_Name_From_Graphing].
Verification Requirements:
Test Scenarios: Define key scenarios to test.
Coverage Points: List critical coverage points.
Test Vectors: Provide 2-3 example test vectors.
Implementation Constraints:
Module-Specific Constraints: Define constraints specific to this module that override or add to the global config .
Optimization Goals: Specify the priority for this block.

Figure 10: Prompts for symbolize.

Prompts for coding. I

System Role Prompt:
You are an expert researcher and software engineer with a deep understanding of experimental design and reproducibility in scientific research.
You will receive a research paper in format, an overview of the plan, a Design in JSON format consisting of "Implementation approach", "File list", "Data structures and interfaces", and "Program call flow", followed by a Task in JSON format that includes "Required packages", "Required other language third-party packages", "Logic Analysis", and "Task list", along with a configuration file named "config.yaml".
Your task is to write code to reproduce the experiments and methodologies described in the paper.
The code you write must be elegant, modular, and maintainable, adhering to Google-style guidelines.
The code must strictly align with the paper's methodology, experimental setup, and evaluation metrics.
Write code with triple quotes.
RTL Coding Prompt:
Based on the planning, design, and tasks of the thesis and the configuration file "config.yaml" specified earlier, write the code in accordance with the "Format Example".
We have done the file list. Next, you must only write the todo file name.
1. Only one file: Do your best to achieve this with only one file.
2. Complete code: Your code will be part of the entire project, so please implement complete, reliable, and reusable code snippets.
3. Follow the design: You must adhere to "Data Structures and Interfaces". Don't change any design. Do not use public member functions that do not exist in the design.
4. Carefully check that no necessary classes/functions are missing in this file.
5. Before using an external variable/module, please make sure to import it first.
6. Write down every detail of the code and do not leave any tasks to be done.
7. Reference configuration: You must use the configuration in "config.yaml". Do not forge any configuration values.

Prompts for coding. II

TB Coding Prompt:
You are an expert who can automate the RTL by using a fully autonomous toolchain. You are invited to teach us how to create chips by using a fully autonomous toolchain for digital layout generation across die sizes, process nodes, and foundry options.
Your specific task now is to write a complete and executable TB for the Verilog module named ‘module_name‘. You must leverage the provided detailed module analysis to infer the module's interface and write a comprehensive and accurate TB.
Here is the detailed analysis of the module and its context from our autonomous toolchain: $module_analysis_content$
In addition, here are a few things you should notice: $rules$

Figure 11: Prompts for coding.

Prompts for Rectifying. I

Error Finding System Role Prompt:
You are an experienced digital circuit engineer, proficient in Verilog/SystemVerilog design, simulation and debugging.
Your task is to analyze the provided Verilog simulation logs, identify the root cause of errors or warnings, and determine which file (such as' rtlfile.v 'or' testbenchfile.v ') is the main responsible for the problem.
Summarize the core lessons learned from this mistake in a concise sentence, with the focus on proposing a universal coding guideline to prevent similar mistakes from happening in the future.

Prompts for Rectifying. III

Error Analysis System Role Prompt:
You are an experienced digital circuit engineer, proficient in Verilog/SystemVerilog design, simulation and debugging.
Your task is to conduct an in-depth analysis of the provided Verilog simulation logs, RTL codes, and TB codes, identify the root causes of errors or warnings, and generate a detailed, fixe-oriented internal analysis result.
This analysis should clearly point out where the problem lies, why it occurs, and the potential methods and considerations for solving it (for example, if it involves width, it is necessary to consider how to correctly crop or expand, or whether the variable definition needs to be modified).
Please output your analysis results in JSON format, including the internal analysis text field.

Prompts for Rectifying. III

Error Fixing System Role Prompt:
You are an experienced digital circuit engineer, proficient in Verilog/SystemVerilog design, simulation and debugging.
Your task is to fix the Verilog code based on the provided original RTL/TBcode, simulation logs, and detailed internal analysis results .
Only modify the necessary code in the problematic files (RTL code and/or TB code) to solve the problem. If a file does not need to be modified, please return its original content. The code should be completely and correctly formatted in the verilog code block.
Please output your reply in JSON format.

Figure 12: Prompts for rectifying.

```
Global Operating Voltage = 1.98
Power-specific unit information :
    Voltage Units = 1V
    Capacitance Units = 1.000000pf
    Time Units = 1ns
    Dynamic Power Units = 1mW    (derived from V,C,T units)
    Leakage Power Units = 1uW


  Cell Internal Power  =   4.1925 mW   (80%)
  Net Switching Power  =   1.0696 mW   (20%)
                         ---------
Total Dynamic Power    =   5.2621 mW  (100%)

Cell Leakage Power     =  69.6676 nW


                 Internal         Switching           Leakage            Total
Power Group      Power            Power               Power              Power   (   %    )  Attrs
--------------------------------------------------------------------------------------------------
io_pad             0.0000           0.0000            0.0000            0.0000  (   0.00%)
memory             0.0000           0.0000            0.0000            0.0000  (   0.00%)
black_box          0.0000           0.0000            0.0000            0.0000  (   0.00%)
clock_network      0.0000           0.0000            0.0000            0.0000  (   0.00%)
register           2.7187           0.1614        1.9938e-02            2.8801  (  54.73%)
sequential         0.0000           0.0000            0.0000            0.0000  (   0.00%)
combinational      1.4738           0.9083        4.9729e-02            2.3821  (  45.27%)
--------------------------------------------------------------------------------------------------
Total              4.1925 mW        1.0696 mW     6.9668e-02 uW        5.2622 mW
```

Figure 13: Reported power of the pe_array.

```
Current design is 'pe_array'.
design_vision> report_area

****************************************
Report : area
Design : pe_array
Version: Q-2019.12-SP5-1
Date   : Thu Jul 31 11:05:32 2025
****************************************

Library(s) Used:

    sc9tap_cm018mg_base_rvt_ff_typical_min_1p98
0-00eac0/arm/tsmc/cm018mg/sc9tap_base_rvt/r0p0/

Number of ports:                          350
Number of nets:                          1409
Number of cells:                          916
Number of combinational cells:            737
Number of sequential cells:               168
Number of macros/black boxes:               0
Number of buf/inv:                         76
Number of references:                       9

Combinational area:               20543.846578
Buf/Inv area:                       532.224007
Noncombinational area:            11802.067291
Macro/Black Box area:                 0.000000
Net Interconnect area:                0.000000

Total cell area:                  32345.913869
Total area:                       32345.913869
```

Figure 15: Reported area of the pe_array.

```
 u_pe0/mult_54/U7/S (ADDHXL)                                  0.22       1.11 f
 u_pe0/mult_54/product[1] (pe_0_DW_mult_uns_0_DW_mult_uns_3)
                                                              0.00       1.11 f
 u_pe0/U29/CO (ADDFX2)                                        0.24       1.35 f
 u_pe0/U28/CO (ADDFX2)                                        0.15       1.50 f
 u_pe0/U23/CO (ADDFX2)                                        0.15       1.66 f
 u_pe0/U21/CO (ADDFX2)                                        0.15       1.81 f
 u_pe0/U18/CO (ADDFX2)                                        0.15       1.96 f
 u_pe0/U15/CO (ADDFX2)                                        0.15       2.11 f
 u_pe0/U14/CO (ADDFX2)                                        0.17       2.28 f
 u_pe0/U129/Y (XOR2X1)                                        0.15       2.43 r
 u_pe0/U22/CO (ADDFX2)                                        0.25       2.68 r
 u_pe0/U20/CO (ADDFX2)                                        0.14       2.82 r
 u_pe0/U19/CO (ADDFX2)                                        0.14       2.96 r
 u_pe0/U12/CO (ADDFX2)                                        0.15       3.11 r
 u_pe0/U110/Y (AND2X1)                                        0.13       3.24 r
 u_pe0/U104/Y (AND2X1)                                        0.13       3.37 r
 u_pe0/U98/Y (AND2X1)                                         0.11       3.49 r
 u_pe0/U95/Y (XOR2X1)                                         0.13       3.61 f
 u_pe0/U13/Y (AOI22X1)                                        0.15       3.76 r
 u_pe0/U16/Y (OAI2BB1X1)                                      0.05       3.81 f
 u_pe0/accum_reg_15_/D (DFFRHQX1)                             0.00       3.81 f
 data arrival time                                                       3.81

 clock clk (rise edge)                                        5.00       5.00
 clock network delay (ideal)                                  0.01       5.01
 clock uncertainty                                           -0.01       5.00
 u_pe0/accum_reg_15_/CK (DFFRHQX1)                            0.00       5.00 r
 library setup time                                          -0.15       4.85
 data required time                                                      4.85
 ------------------------------------------------------------------------------
 data required time                                                      4.85
 data arrival time                                                      -3.81
 ------------------------------------------------------------------------------
 slack (MET)                                                             1.04
```

Figure 14: Reported performance of the pe_array.

```
Global Operating Voltage = 1.98
Power-specific unit information :
    Voltage Units = 1V
    Capacitance Units = 1.000000pf
    Time Units = 1ns
    Dynamic Power Units = 1mW    (derived from V,C,T units)
    Leakage Power Units = 1uW


  Cell Internal Power  = 273.2457 uW   (84%)
  Net Switching Power  =  51.1611 uW   (16%)
                         ---------
Total Dynamic Power    = 324.4068 uW  (100%)

Cell Leakage Power     =   2.6998 nW


                 Internal         Switching           Leakage            Total
Power Group      Power            Power               Power              Power   (   %    )  Attrs
--------------------------------------------------------------------------------------------------
io_pad             0.0000           0.0000            0.0000            0.0000  (   0.00%)
memory             0.0000           0.0000            0.0000            0.0000  (   0.00%)
black_box          0.0000           0.0000            0.0000            0.0000  (   0.00%)
clock_network      0.0000           0.0000            0.0000            0.0000  (   0.00%)
register           0.2563        1.1524e-02        1.9077e-03            0.2679  (  82.57%)
sequential         0.0000           0.0000            0.0000            0.0000  (   0.00%)
combinational  1.6905e-02        3.9637e-02        7.9204e-04        5.6543e-02  (  17.43%)
--------------------------------------------------------------------------------------------------
Total              0.2732 mW     5.1161e-02 mW     2.6998e-03 uW        0.3244 mW
```

Figure 16: Reported power of the spark_encoder.

```
          (rising edge-triggered flip-flop clocked by clk)
  Path Group: clk
  Path Type: max

  Des/Clust/Port     Wire Load Model       Library
  ------------------------------------------------
  spark_encoder      Small                 sc9tap_cm018mg_base

  Point                                    Incr       Path
  -----------------------------------------------------------
  clock clk (rise edge)                    0.00       0.00
  clock network delay (ideal)              0.01       0.01
  input external delay                     0.70       0.71 f
  valid_in (in)                            0.00       0.71 f
  U32/Y (NAND2X1)                          0.23       0.94 r
  U47/Y (INVX1)                            0.24       1.18 f
  U26/Y (AOI22X1)                          0.15       1.34 r
  U25/Y (INVX1)                            0.04       1.38 f
  raw_reg_reg_5_/D (DFFRHQX1)              0.00       1.38 f
  data arrival time                                   1.38

  clock clk (rise edge)                    5.00       5.00
  clock network delay (ideal)              0.01       5.01
  clock uncertainty                       -0.01       5.00
  raw_reg_reg_5_/CK (DFFRHQX1)             0.00       5.00 r
  library setup time                      -0.15       4.85
  data required time                                  4.85
  -----------------------------------------------------------
  data required time                                  4.85
  data arrival time                                  -1.38
  -----------------------------------------------------------
  slack (MET)                                         3.47
```

Figure 17: Reported performance of the spark_encoder.

```
****************************************
Report : area
Design : spark_encoder
Version: Q-2019.12-SP5-1
Date   : Thu Jul 31 10:38:12 2025
****************************************

Library(s) Used:

    sc9tap_cm018mg_base_rvt_ff_typical_min_1p98v_0c
FB-00000-r0p0-00eac0/arm/tsmc/cm018mg/sc9tap_base_r
_1p98v_0c.db)

Number of ports:                            17
Number of nets:                             71
Number of cells:                            60
Number of combinational cells:              44
Number of sequential cells:                 16
Number of macros/black boxes:                0
Number of buf/inv:                          17
Number of references:                       15

Combinational area:                 558.835208
Buf/Inv area:                       113.097601
Noncombinational area:             1117.670410
Macro/Black Box area:                 0.000000
Net Interconnect area:                0.000000

Total cell area:                   1676.505619
Total area:                        1676.505619
1
design_vision> report_clock
```

Figure 18: Reported area of the spark_encoder.

```
****************************************
Report : area
Design : spark_decoder
Version: Q-2019.12-SP5-1
Date   : Thu Jul 31 10:59:38 2025
****************************************

Library(s) Used:

    sc9tap_cm018mg_base_rvt_ff_typical_min_1p98
0-00eac0/arm/tsmc/cm018mg/sc9tap_base_rvt/r0p0/

Number of ports:                            12
Number of nets:                             25
Number of cells:                            18
Number of combinational cells:              12
Number of sequential cells:                  6
Number of macros/black boxes:                0
Number of buf/inv:                           6
Number of references:                        6

Combinational area:                 149.688001
Buf/Inv area:                        39.916800
Noncombinational area:              419.126404
Macro/Black Box area:                 0.000000
Net Interconnect area:                0.000000

Total cell area:                    568.814404
Total area:                         568.814404
```

Figure 19: Reported power of the spark_decoder.

```
  Path Group: clk
  Path Type: max

  Des/Clust/Port     Wire Load Model       Library
  ------------------------------------------------
  spark_decoder      Small                 sc9tap_cm018mg_base_

  Point                                    Incr       Path
  -----------------------------------------------------------
  clock clk (rise edge)                    0.00       0.00
  clock network delay (ideal)              0.01       0.01
  input external delay                     0.70       0.71 r
  in_valid (in)                            0.00       0.71 r
  U13/Y (INVX1)                            0.09       0.80 f
  U12/Y (AOI32X1)                          0.13       0.93 r
  U11/Y (INVX1)                            0.04       0.98 f
  post_flag_reg/D (DFFRHQX1)               0.00       0.98 f
  data arrival time                                   0.98

  clock clk (rise edge)                    5.00       5.00
  clock network delay (ideal)              0.01       5.01
  clock uncertainty                       -0.01       5.00
  post_flag_reg/CK (DFFRHQX1)              0.00       5.00 r
  library setup time                      -0.15       4.85
  data required time                                  4.85
  -----------------------------------------------------------
  data required time                                  4.85
  data arrival time                                  -0.98
  -----------------------------------------------------------
  slack (MET)                                         3.88
```

Figure 20: Reported performance of the spark_decoder.

```
spark_decoder          Small             sc9tap_cm018mg_base_rvt_ff_typical_min_1p98v_0c


Global Operating Voltage = 1.98
Power-specific unit information :
    Voltage Units = 1V
    Capacitance Units = 1.000000pf
    Time Units = 1ns
    Dynamic Power Units = 1mW    (derived from V,C,T units)
    Leakage Power Units = 1uW


  Cell Internal Power  = 100.0763 uW   (92%)
  Net Switching Power  =   8.9223 uW    (8%)
                         ---------
Total Dynamic Power    = 108.9985 uW  (100%)

Cell Leakage Power     = 938.0424 pW


                 Internal         Switching           Leakage            Total
Power Group      Power            Power               Power              Power   (   %    )  Attrs
--------------------------------------------------------------------------------------------------
io_pad             0.0000           0.0000            0.0000            0.0000  (   0.00%)
memory             0.0000           0.0000            0.0000            0.0000  (   0.00%)
black_box          0.0000           0.0000            0.0000            0.0000  (   0.00%)
clock_network      0.0000           0.0000            0.0000            0.0000  (   0.00%)
register       9.5663e-02       2.8983e-03        7.2920e-04        9.8562e-02  (  90.42%)
sequential         0.0000           0.0000            0.0000            0.0000  (   0.00%)
combinational  4.4135e-03       6.0239e-03        2.0884e-04        1.0438e-02  (   9.58%)
--------------------------------------------------------------------------------------------------
Total              0.1001 mW     8.9223e-03 mW     9.3804e-04 uW        0.1090 mW
```

Figure 21: Reported area of the spark_decoder.