LOCALV: EXPLOITING INFORMATION LOCALITY FOR IP-LEVEL VERILOG GENERATION

Anonymous authors
Paper under double-blind review

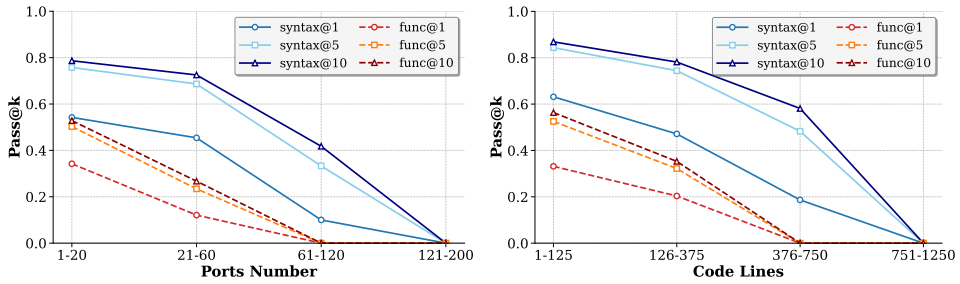
ABSTRACT

The generation of Register-Transfer Level (RTL) code is a crucial yet labor-intensive step in digital hardware design, traditionally requiring engineers to manually translate complex specifications into thousands of lines of synthesizable Hardware Description Language (HDL) code. While Large Language Models (LLMs) have shown promise in automating this process, existing approaches—including fine-tuned domain-specific models and advanced agent-based systems—struggle to scale to industrial IP-level design tasks. We identify three key challenges: (1) handling long, highly detailed documents, where critical interface constraints become buried in unrelated submodule descriptions; (2) generating long RTL code, where both syntactic and semantic correctness degrade sharply with increasing output length; and (3) navigating the complex debugging cycles required for functional verification through simulation and waveform analysis. To overcome these challenges, we propose *LocalV*, a multi-agent framework that leverages the inherent *information locality* in modular hardware design. LocalV decomposes the long-document to long-code generation problem into a set of short-document, short-code tasks, enabling scalable generation and debugging. Specifically, LocalV integrates hierarchical document partitioning, task planning, localized code generation, interface-consistent merging, and AST-guided locality-aware debugging. Experiments on REALBENCH demonstrate that LocalV substantially outperforms state-of-the-art (SOTA) LLMs and agents, showing the potential of generating Verilog for IP-level RTL design.

1 INTRODUCTION

The generation of Register-Transfer Level (RTL) code is a core step in digital hardware design. This process is notoriously labor-intensive and error-prone, as engineers must manually translate natural language specifications into thousands of lines of synthesizable Hardware Description Language (HDL) code (e.g., Verilog, VHDL). The promise of Large Language Models (LLMs) to automate this step has spurred rapid innovation. Initial efforts focused on benchmarking general-purpose models (Liu et al., 2023b; Thakur et al., 2023) and developing domain-specific solutions through fine-tuning or data augmentation (Liu et al., 2024c; Cui et al., 2024; Liu et al., 2024b; Zhao et al., 2025). More recently, the field has shifted towards sophisticated agent-based systems that mimic human design workflows. These agents, such as VerilogCoder (Ho et al., 2025) and MAGE (Zhao et al., 2024), decompose complex problems and can operate autonomously or in a human-in-the-loop fashion, as explored in collaborative design platforms like ChatCPU (Wang et al., 2024) and Spec2RTL-Agent (Yu et al., 2025).

Despite strong results on academic benchmarks like VerilogEval (Liu et al., 2023b), a clear gap appears when applying current LLM-based methods to industrial hardware design. This is particularly evident with REALBENCH (Jin et al., 2025), an IP-level benchmark derived from real-world open-source IP, which features significantly longer documentation (197.3 vs. 5.7) and code lengths (241.2 vs. 15.8) compared to VerilogEval. Directly using SOTA models or agents often leads to a sharp drop in performance, with many outputs failing to be even syntactically correct, let alone functionally valid. This gap highlights a mismatch between current model capabilities and the high requirements of real-world hardware engineering. We observe three main challenges:



(a) Pass@k vs. I/O signal count

(b) Pass@k vs. code length

Figure 1: Performance of Claude 3.7 Sonnet on REALBENCH: Pass@k vs. (a) I/O signal count and (b) code length (lines), reporting syntactic and functional Pass@k. Accuracy decreases with interface complexity and output length.

Long-Document Handling. IP-level hardware specifications are typically verbose and detailed, largely due to the increasing number of I/O signals and submodules. Although modern LLMs support context windows of 32k tokens or more, their ability to generate functionally correct RTL code diminishes as document complexity grows. The accumulation of signal and module details overwhelms the model's limited understanding of hardware semantics. As a result, critical interface constraints are often obscured by irrelevant details, leading to phantom signals, port list mismatches, and logically incorrect Verilog code. This trend is illustrated in Figure 1a, which shows a consistent decrease in LLM accuracy as the number of I/O signals increases.

Long-Code Generation. IP-level designs usually involve substantially longer code, which exacerbates the challenges LLMs face in HDL generation—a domain where they already underperform. As shown in Figure 1b, both syntactic and semantic accuracy drop significantly with code length. When the code exceeds 750 lines, even repeated sampling (e.g., 10 times) fails to yield a syntactically valid result. Typical errors include incorrect macro references, use of non-synthesizable constructs, and fundamental syntax errors, underscoring the model's inherent limitations in generating reliable RTL code.

Complex Debugging Process. In practice, IP-level hardware verification relies on carefully constructed testbenches to ensure the design conforms to specifications. Each simulation failure triggers a laborious debugging cycle: engineers analyze waveforms to identify faulty signals, trace errors back to ambiguous or misinterpreted specification segments, and iteratively refine the design. This process not only corrects the code but also clarifies ambiguities in the specification itself, using waveform behavior as a definitive reference for refinement.

To address these challenges, we propose LocalV, a multi-agent framework explicitly designed for the real-world IP-level "long-document, long-code" hardware generation problem. Our key observation is that IP-level specifications inherit strong **information locality** from modular hardware design: code fragments can often be generated correctly by relying on only a portion of the document. This suggests that long-document to long-code generation can be decomposed into a set of short-document to short-code tasks without information loss, thereby mitigating the core challenges.

Specifically, LocalV organizes the following workflow as shown in Figure 2: (1) Preprocessing. Documents are partitioned into fragments with hierarchical indices. (2) Planning. Code structure is planned as sub-tasks with assigned document fragments. (3) Generation. Coding agents execute "short-document, short-code" generation for each sub-task. (4) Merging. Fragments are merged into a complete design with interface consistency. (5) Debugging. Error messages and AST-guided waveform analysis trace failures back to specification fragments for locality-aware debugging.

Our contributions are summarized as follows: (1) We realized three fundamental challenges in generating IP-level Verilog code, namely, long-document handling, long-code generation, and the complex debugging process. (2) We observed the information locality principle for IP-level Verilog generation. Based on it, we introduce an index-driven document partitioning mechanism, a fragment-based generation strategy that decomposes complex tasks into manageable subtasks, and a traceable debugging pipeline that maps errors back to relevant specification fragments via AST-guided analysis. (3) Based on these techniques, we present LocalV, a multi-agent framework for

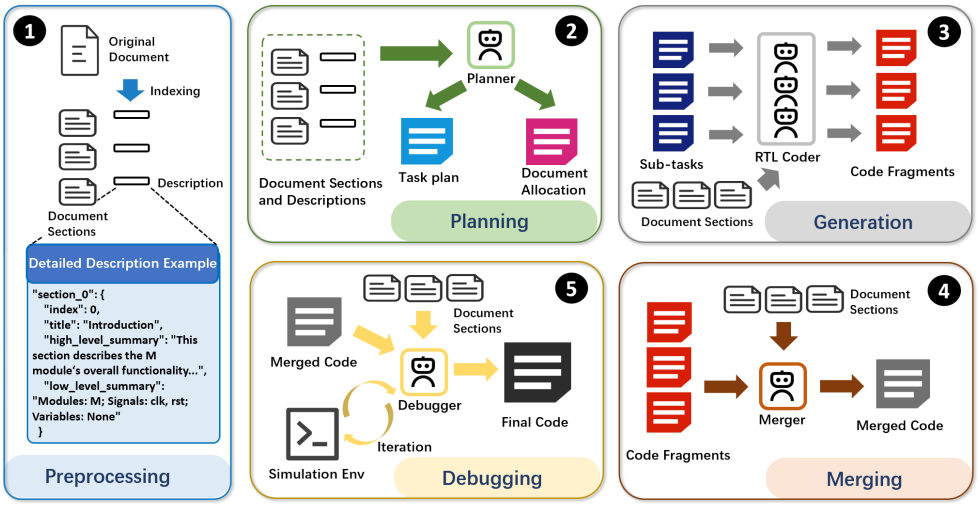


Figure 2: Workflow overview of LocalV.

generating correct Verilog from **IP-level specifications**. LocalV achieves a 45.0% pass rate on REALBENCH (Jin et al., 2025), surpassing SOTA agent-based frameworks by 23.4%.

2 Related Work

LLM-based RTL Generation. The application of LLMs to automate RTL code generation has emerged as a promising research area in electronic design automation (EDA). Early explorations (Nair et al., 2023; Chang et al., 2023; Blocklove et al., 2023) focused on evaluating the capability of general-purpose LLMs to translate natural language specifications into Hardware Description Languages (HDLs) like Verilog and VHDL. Foundational benchmarks such as VerilogEval (Liu et al., 2023b) and RTLLM (Lu et al., 2024) were established to systematically assess model performance, revealing both the potential and limitations of off-the-shelf models. To improve performance, subsequent research has focused on domain-specific adaptation through fine-tuning on curated datasets (Liu et al., 2024c; Thakur et al., 2024; Liu et al., 2023a) or optimization via reinforcement learning (Pei et al., 2024; Zhu et al., 2025; Chen et al., 2025). While these models show strong results on well-defined, smaller-scale problems, their effectiveness on real-world, IP-level specifications is fundamentally limited. For many fine-tuned models, this stems from smaller model scales, the lack of training data for IP-level hardware design, and constrained context windows that fail to fully capture complex design documents. More critically, even for large-scale models with extensive context capabilities, the single-pass generation paradigm is ill-suited for the complexity of IP-level design. Attempting to synthesize functionally correct code from a verbose specification in a single attempt struggles to capture the intricate dependencies and hierarchical nature of hardware, often leading to subtle but critical errors.

Agent-based Frameworks for Hardware Design. To overcome the limitations of single-pass generation, the field is shifting towards multi-agent frameworks that emulate the collaborative and iterative nature of human design and verification workflows. This paradigm moves beyond a single monolithic model to a team of specialized agents, each assigned a distinct role. For instance, MAGE (Zhao et al., 2024) explicitly creates a four-agent team responsible for RTL generation, testbench creation, functional evaluation (judging), and debugging, establishing a clear, recursive loop of proposing and refining the design. Similarly, RTLSquad (Wang et al., 2025) organizes its agents into ”squads” dedicated to distinct project phases—exploration, implementation, and verification—thereby mimicking the structure of a human engineering team. Central to these systems is a task decomposition phase, where a high-level specification is broken down into a structured plan with manageable sub-tasks. These plans guide the execution of agents focused on coding, planning, and reflection, as seen in frameworks like Spec2RTL-Agent (Yu et al., 2025) and VerilogCoder (Ho et al., 2025). However, this decomposition process faces a critical challenge: translating the orig-

inal specification into intermediate instructions can introduce cascading ambiguity, distorting the design intent. Consequently, debugging becomes severely hampered, as agents must trace errors through these distorted interpretations rather than the source document. Our approach addresses this by maintaining a direct link between the specification and code, grounding the entire process in the original document.

3 METHODOLOGY

We begin by formalizing the problem of IP-level Verilog generation (§3.1). We then introduce our core hypothesis, the *information locality* in IP-level hardware specifications, with a quantitative analysis (§3.2). We then present the detailed LocalV pipeline built on this insight(§3.3).

3.1 PROBLEM FORMULATION

Our objective is to synthesize a complete Verilog module from a natural language specification. We formally define the problem as follows:

Input: A natural language specification document $\mathcal{D}$, represented as an ordered sequence of N semantic textual units (e.g., paragraphs or sections), $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$. Also, a target module name m and a simulation environment E that provides golden execution feedback (including error messages and behavioral mismatches) for debugging purposes is given.

Output: A synthesizable Verilog module $\mathcal{V}_m$. We model the generated code not as a monolithic text file, but as a structured set of M semantic code units, $\mathcal{V}_m = \{c_1, c_2, \ldots, c_M\}$. A code unit c_j represents a functionally cohesive and syntactically complete block of RTL code, such as a module or a statement. The final output file is the concatenation of these units.

Objective: The generated module $\mathcal{V}_m$ must be functionally correct and can pass a suite of simulation tests from E against a golden reference testbench, ensuring functional equivalence.

3.2 INFORMATION LOCALITY

Our approach is grounded in a core assumption we term **Information Locality**: for any semantic code unit $c_j \in \mathcal{V}_m$, the information required to generate c_j is primarily concentrated within a subset of the specification $\mathcal{D}$. This locality arises directly from the hierarchical and modular nature of hardware design. Complex systems are built from well-defined submodules (e.g., ALUs, register files), and IP-level specifications explicitly mirror this structure: dedicated sections describe each module's behavior, I/O, and internal logic. This creates a natural alignment, where the implementation of a code unit c_j depends predominantly on its corresponding documentation segment. In contrast, general-purpose software specifications often describe high-level algorithms that do not decompose neatly into code-level constructs, leading to more diffuse information sources.

The validity of this locality hypothesis is key. If it holds—as we will quantitatively demonstrate below (Figure 3)—it enables a powerful divide-and-conquer strategy. The quantitative results confirm that the "long-document to long-code" mapping problem can be effectively decomposed into a set of parallelizable "short-document to short-code" subproblems without information loss, dramatically improving tractability.

We quantify information locality by measuring the entropy of the information source distribution for each code unit. Our analysis begins by segmenting the specification $\mathcal{D}$ into paragraphs $\{d_i\}_{i=1}^{N}$ and the Verilog code $\mathcal{V}_m$ into statements $\{c_j\}_{j=1}^{M}$. For each code statement c_j, we compute its semantic similarity $\mathrm{sim}(d_i, c_j)$ (cosine similarity of Qwen3-Embedding-0.6B embeddings) to every specification paragraph d_i and transform them into conditional probability distribution $P(d_i \mid c_j)$ using a softmax function with temperature $\tau = 0.1$:

$$P(d_i \mid c_j) = \frac{\exp(\mathrm{sim}(d_i, c_j)/\tau)}{\sum_{k=1}^{N} \exp(\mathrm{sim}(d_k, c_j)/\tau)}. \tag{1}$$

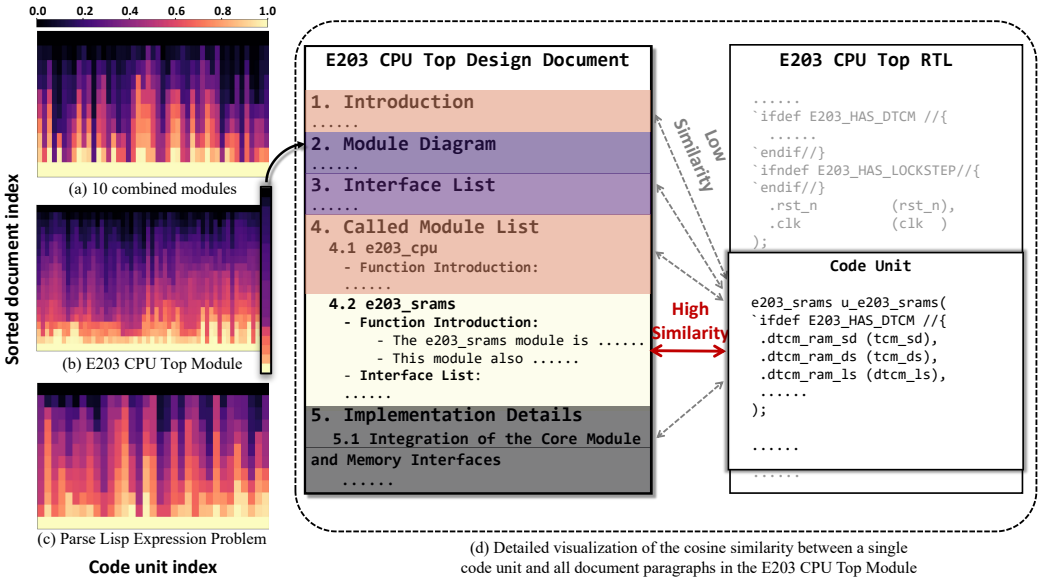


Figure 3: Heatmaps of normalized cosine similarity across three tasks. Each column represents a code unit and its sorted cosine similarity to all document paragraphs. Values in each column are independently normalized to the range [0, 1], where lower values indicate higher information locality. (a) 10 randomly selected and then combined modules from VerilogEval, demonstrating extremely high information locality (since they are totally independent of each other) with $\bar{H}_{\text{norm}} = 0.8206$. (b) The E203 CPU Top Module from REALBENCH, showing high information locality with $\bar{H}_{\text{norm}} = 0.8680$. (c) The Parse Lisp Expression problem, a typical software task, with $\bar{H}_{\text{norm}} = 0.9126$. (d) A detailed visualization of the cosine similarity between a single code unit and all document paragraphs in the E203 CPU Top Module.

The locality for c_j is then assessed by the entropy of this distribution:

$$H(c_j) = -\sum_{i=1}^{N} P(d_i \mid c_j) \log_2 P(d_i \mid c_j), \tag{2}$$

where lower entropy indicates that information is concentrated in a small number of textual units, thus supporting the locality hypothesis.

To ensure comparability across specifications of different lengths, we normalize the entropy by its theoretical maximum, $H_{\max} = \log_2 N$, which occurs under a uniform distribution. The normalized entropy for a code unit is:

$$H_{\text{norm}}(c_j) = \frac{H(c_j)}{\log_2 N}. \tag{3}$$

This yields a scale-invariant measure where $H_{\text{norm}}(c_j) \in [0, 1]$. To report a single locality score for an entire design, we average the normalized entropy across all M code units:

$$\bar{H}_{\text{norm}} = \frac{1}{M} \sum_{j=1}^{M} H_{\text{norm}}(c_j). \tag{4}$$

A lower $\bar{H}_{\text{norm}}$ indicates stronger overall locality, and this metric is comparable across experiments with varying N and M.

We evaluate three settings to contrast information locality: (a) a **synthetic Verilog benchmark** (10 concatenated VerilogEval cases) as an ideal locality baseline (lower bound); (b) the **hardware IP** e203_cpu_top from REALBENCH; and (c) a **software counterpart** (LeetCode "Parse Lisp Expression" in Python) with comparable length. Row-normalized heatmaps and the average normalized entropy $\bar{H}_{\text{norm}}$ quantify locality strength. As shown in Figure 3, the hardware design (b) exhibits strong locality ($\bar{H}_{\text{norm}} = 0.8680$), much closer to the ideal (a) ($\bar{H}_{\text{norm}} = 0.8206$) than the software case (c) ($\bar{H}_{\text{norm}} = 0.9126$). This pattern holds across REALBENCH, where the average $\bar{H}\text{norm} = 0.8406$ further confirms stronger locality in hardware specifications.

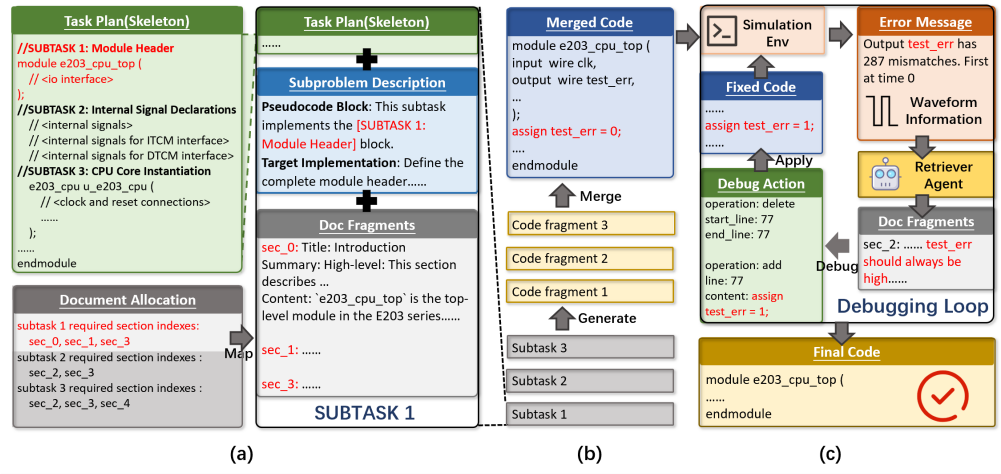


Figure 4: The detailed workflow of LocalV. (a) Output of the planning stage, illustrating the structure of a sub-task. (b) Overview of the code generation and merging process. (c) Overview of the debugging loop and the generation of the final code.

3.3 LocalV Overview

We now introduce our novel multi-agent framework, **LocalV**, designed to automate the generation of Verilog code from long natural language documentation. The overall workflow is depicted in Figure 4 and detailed in the subsequent sections.

3.3.1 Preprocessing

The first stage of our pipeline structures the input documentation for efficient retrieval and comprehension by the agents. Given a raw design document, we split the text into coherent paragraphs. For each paragraph, an LLM is prompted to generate a dual-level description that indexes the source content:

Semantic level: Provides a high-level summary of the paragraph's functional intent, such as "interface specification for the DMA controller" or "timing constraints for the DDR memory interface." This supports agents who require a conceptual understanding of a module's purpose.

Lexical level: Extracts fine-grained hardware-specific entities—including signal names, module identifiers, macros, and parameters—to ensure precise retrieval of low-level details that may be omitted in semantic summaries.

The resulting description serves as keys indexing the original text segments and is used in subsequent stages of the pipeline.

3.3.2 Planning and Task Decomposition

Upon receiving the indexed documentation from the previous stage, the **Planner Agent** constructs the overall structure of the final Verilog code and generates a corresponding skeleton. This skeleton is expressed as pseudo-code containing syntactic placeholders that represent various code components—such as submodule instantiations or signal assignments.

The agent then decomposes the skeleton into sub-tasks, each corresponding to a code fragment that requires implementation. For every sub-task, the **Retriever** queries the hierarchical index to identify and retrieve the most relevant document sections, attaching them as focused context. This targeted contextualization not only narrows the scope of each generation step but also ensures alignment with the original specification.

Unlike approaches that naively partition hardware into submodules or create intermediate representations, our fragment-based decomposition introduces no additional complexity. All sub-tasks contribute directly to the same global design, each addressing a well-defined portion of the code.

This method maintains tight alignment with the final output and mitigates common issues such as objective drift that may arise from self-generated intermediate goals.

3.3.3 RTL GENERATION

With the sub-tasks and their associated documentation contexts prepared, multiple instances of the **RTL Agent** proceed to fill the placeholders in the code skeleton. Each agent is assigned a specific sub-task and operates within a constrained context, allowing it to focus exclusively on its local objective. This narrow focus facilitates an accurate translation of the specification into synthesizable Verilog for the corresponding code segment, thereby reducing errors such as phantom signals and enhancing the overall quality of the generated code fragments.

3.3.4 CODE FRAGMENTS MERGING

After all **RTL Agents** complete fragment generation, the **Merge Agent** integrates the fragments into a correct Verilog module. To resolve potential inconsistencies or implementation errors that may arise during merging, the **Retriever Agent** first fetches relevant sections from the original documentation. Using this retrieved context, the **Merge Agent** then refines and integrates the fragments using this additional information together with the generated code, ensuring that the final output is correct and coherent.

3.3.5 LOCALITY-AWARE DEBUGGING

LocalV's debugging pipeline leverages **information locality** to efficiently trace errors back to their relevant documentation segments. The process begins by curating error messages from the simulation environment to extract key signals—such as syntax error locations or functional mismatches, and root-cause signal information from waveform analysis (inspired by VerilogCoder's (Ho et al., 2025)). Crucially, the **Retriever Agent** then uses this error context to fetch the small subset of documentation fragments that are locally relevant to the faulty code section, as determined by the underlying information locality hypothesis. A dedicated **Debug Agent** subsequently synthesizes this focused context—the error details and the retrieved documentation—to produce precise, line-number-aware edit actions (e.g., inserting or deleting specific lines). This debug loop iterates until the code is error-free or a predefined iteration limit is reached, efficiently minimizing corrective overhead by avoiding reprocessing the entire specification.

4 EXPERIMENTS

We evaluate LocalV's performance on realistic hardware design tasks through a series of experiments. We first describe our experimental setup, then report the main results comparing LocalV against baselines, and finally perform an ablation study to quantify the contribution of components.

4.1 SETTINGS

Benchmarks. We adopt REALBENCH (Jin et al., 2025), a challenging benchmark specifically designed for real-world, IP-level Verilog generation. REALBENCH comprises 60 RTL generation tasks drawn from three IPs: AES encoder/decoder cores (6 modules), an SD card controller (14 modules), and a CPU core (40 modules). REALBENCH emphasizes practical applicability through long-form natural language specifications (averaging 10k tokens) and substantial implementation complexity (approximately 320 lines of Verilog code per target module on average).

Metrics. We evaluate models on syntactic and functional correctness using REALBENCH's predefined testbenches, and report both syntax and functional pass rate as the metric. We use Pass@1 in Table 1 2, and extend to Pass@k (OpenAI, 2021; Liu et al., 2023b) in some analysis. As reported in Table 1, the pass rates for direct prompting model baselines are averaged over 20 independent generations per task, whereas Agent baselines are evaluated using a single generation per task.

Baselines. We establish comprehensive baselines comprising both standalone models and agent-based systems. For standalone models, we evaluate both commercial and open-source mod-

Table 1: Syntax and functional ppass rate comparison on the REALBENCH benchmark.

	SDC		AES		E203 CPU		ALL	
Method	Syn.	Func.	Syn.	Func.	Syn.	Func.	Syn.	Func.
Model Baselines								
Claude-3.7	41.4%	11.7%	46.6%	31.6%	42.7%	20.6%	42.8%	19.6%
DeepSeek-V3	44.2%	15.3%	55.8%	23.3%	19.5%	7.5%	28.9%	10.9%
DeepSeek-R1	28.5%	7.1%	66.6%	**50.0%**	12.5%	10.0%	21.6%	13.3%
Qwen3-32B	25.3%	15.3%	32.4%	16.6%	8.3%	6.2%	14.7%	9.4%
GPT-4o	14.2%	0.0%	50.0%	16.6%	5.0%	0.0%	11.6%	1.6%
GPT-5	7.1%	0.0%	50.0%	33.3%	30.0%	20.0%	26.6%	16.6%
Agent Baselines								
MAGE (Claude)	57.1%	21.4%	66.6%	33.3%	62.5%	20.0%	61.6%	21.6%
VerilogCoder (Claude)	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
LocalV (DeepSeek-V3)	64.2%	28.5%	50.0%	**50.0%**	60.0%	35.0%	60.0%	35.0%
LocalV (Claude)	**78.5%**	**35.7%**	**83.3%**	**50.0%**	**72.5%**	**47.5%**	**75.0%**	**45.0%**

els, including Claude (Claude-3.7-sonnet-250219) (Anthropic, 2025), DeepSeek-V3 (DeepSeek-v3-250324) (Liu et al., 2024a), DeepSeek-R1 (DeepSeek-r1-250528) (Guo et al., 2025), Qwen3-32B (Yang et al., 2025), GPT-4o (OpenAI, 2024), and GPT-5 (OpenAI, 2025). For agent-based approaches, we compare against SOTA methods, MAGE (Zhao et al., 2024) and VerilogCoder (Ho et al., 2025), both implemented using Claude-3.7-sonnet-250219. Our LocalV method is evaluated on two different backbone models: Claude-3.7-sonnet-250219 and DeepSeek-v3-250324.

4.2 MAIN RESULTS

Table 1 presents the evaluation results on the challenging REALBENCH benchmark. This benchmark proves particularly difficult for current LLMs, as evidenced by the modest 19.0% functional Pass@1 achieved even by the strong Claude-3.7-sonnet-250219 model. Notably, our LocalV (Claude) surpasses the base model's Pass@20 performance (35.0%) with just a single generation, highlighting its significant advantages for IP-level hardware design tasks.

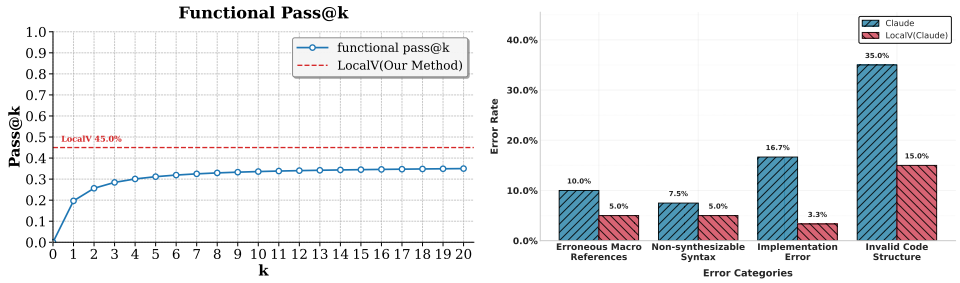
When compared against agent baselines, LocalV demonstrates superior performance over both MAGE (overall 23.4%) and VerilogCoder. It is important to note that MAGE typically relies on extensive high-temperature sampling to generate candidate programs—a computationally expensive approach for long-form code generation. To ensure a fair comparison with LocalV's single-shot setting, we limited MAGE's candidate size to two and allocated an equivalent debugging iteration budget. VerilogCoder employs a ReAct-style workflow (Yao et al., 2023) that performs well on simpler tasks but struggles with IP-level complexity. Without specific design adaptations for complex hardware generation, its per-agent success rates diminish as context length increases, and its nondeterministic orchestration leads to high computational costs and low completion rates. Under reasonable cost constraints, VerilogCoder failed to solve any REALBENCH instances.

In contrast, LocalV achieves stronger performance with substantially improved resource efficiency, enabled by its streamlined agent architecture and precise task decomposition strategy. The method generates each code fragment only once, performs a single merge operation, and executes a bounded debugging schedule (maximum 10 iterations), producing high-quality solutions while maintaining controlled generation costs.

Also, we present a comparison between LocalV and direct sampling using Claude in Figure 5. To demonstrate the superior functional accuracy of LocalV, we plot its performance alongside the Pass@k values of direct sampling in Figure 5a. The results indicate that the Pass@k of direct sampling tends to converge after $k = 10$, yet remains substantially lower than the accuracy achieved by LocalV. Furthermore, we provide a detailed breakdown of syntax error categories for both methods in Figure 5b. The results show that LocalV consistently exhibits a lower syntax error rate across all categories, highlighting its robust syntactic performance in diverse problem settings.

4.3 ABLATION STUDY

We conduct ablation studies on LocalV (based on Claude) in Table 2.



(a) Sampling results of Claude 3.7 Sonnet vs. LocalV.

(b) Distribution of syntactic error types for Claude 3.7 Sonnet and LocalV.

Figure 5: Comparison between LocalV and direct sampling

Table 2: Ablation study on the REALBENCH benchmark.

	SDC		AES		E203 CPU		ALL	
Method	Syn.	Func.	Syn.	Func.	Syn.	Func.	Syn.	Func.
LocalV	**78.5%**	**35.7%**	**83.3%**	**50.0%**	**72.5%**	**47.5%**	**75.0%**	**45.0%**
w/o index	64.2%	21.4%	100.0%	**50.0%**	57.5%	37.5%	63.3%	35.0%
w/o index & debug	35.7%	7.1%	50.0%	33.3%	57.5%	22.5%	51.6%	20.0%
w/o index & debug & plan	35.7%	7.1%	50.0%	33.3%	45.0%	22.5%	43.3%	20.0%

First, replacing indexed document fragments with the full specification significantly degrades performance across all benchmarks. The hierarchical indexing mechanism proves essential for managing IP-level complexity, as long specifications introduce substantial irrelevant content that distracts the model and harms generation quality. Even with other components intact, removing indexing alone causes a notable 10.0% drop in overall functional pass rate.

Second, the debugging component demonstrates the importance of code correctness. When both indexing and debugging are removed, performance drops to 20.0%—only marginally above the base model's Pass@1. This indicates that while our task decomposition strategy addresses core challenges, the debugging stage is vital for ensuring functional correctness of the generated IP blocks.

Finally, the planner provides complementary benefits by enhancing syntactic correctness and orchestrating the generation process. While its impact on functional accuracy is less pronounced than indexing and debugging, it contributes to syntactic accuracy, and the full ablation (without index, planner, and debug) yields the lowest performance (20.0%), confirming the planner's role in maintaining structural coherence for complex hardware design tasks.

Overall, these results confirm that information locality is the unifying principle behind LocalV's effectiveness. The hierarchical indexing establishes locality by focusing on relevant document fragments, while the planner maintains locality through structured generation. The debugging component extends this approach by tracing errors to specific documentation segments for targeted corrections. The performance degradation when compromising locality—whether through fragmented generation without indexing or monolithic generation without planning—demonstrates that locality-aware decomposition is essential for IP-level code generation under constrained budgets.

5 CONCLUSION

We present LOCALV, a multi-agent framework with a workflow tailored to IP-level hardware design. Our study observes and validates the information locality of IP-level hardware specifications. Most RTL fragments can be correctly implemented based on a partial specification. Building on this insight, we design a novel hierarchical indexing strategy, a fragment-oriented task decomposition, and a locality-aware debugging loop. In REALBENCH, a real-world IP-level benchmark, LocalV delivers a 10% improvement, advancing the practical generation of reliable RTL code with LLM.

REFERENCES

Anthropic. Claude 3.7 sonnet, Feb 2025. URL https://www.anthropic.com/news/claude-3-7-sonnet.

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

Kaiyan Chang, Ying Wang, Haimeng Ren, Mengdi Wang, Shengwen Liang, Yinhe Han, Huawei Li, and Xiaowei Li. Chipgpt: How far are we from natural language hardware design. *arXiv preprint arXiv:2305.14019*, 2023.

Zhirong Chen, Kaiyan Chang, Zhuolin Li, Xinyang He, Chujie Chen, Cangyuan Li, Mengdi Wang, Haobo Xu, Yinhe Han, and Ying Wang. Chipseek-r1: Generating human-surpassing rtl with llm via hierarchical reward-driven reinforcement learning. *arXiv preprint arXiv:2507.04736*, 2025.

Fan Cui, Chenyang Yin, Kexing Zhou, Youwei Xiao, Guangyu Sun, Qiang Xu, Qipeng Guo, Yun Liang, Xingcheng Zhang, Demin Song, et al. Origen: Enhancing rtl code generation with code-to-code augmentation and self-reflection. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pp. 1–9, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chia-Tung Ho, Haoxing Ren, and Brucek Khailany. 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, pp. 300–307, 2025.

Pengwei Jin, Di Huang, Chongxiao Li, Shuyao Cheng, Yang Zhao, Xinyao Zheng, Jiaguo Zhu, Shuyi Xing, Bohan Dou, Rui Zhang, et al. Realbench: Benchmarking verilog generation models with real-world ip designs. *arXiv preprint arXiv:2507.16200*, 2025.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, et al. Chipnemo: Domain-adapted llms for chip design. *arXiv preprint arXiv:2311.00176*, 2023a.

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

Mingjie Liu, Yun-Da Tsai, Wenfei Zhou, and Haoxing Ren. Craftrtl: High-quality synthetic data generation for verilog code models with correct-by-construction non-textual representations and targeted code repair. *arXiv preprint arXiv:2409.12993*, 2024b.

Shang Liu, Wenji Fang, Yao Lu, Jing Wang, Qijun Zhang, Hongce Zhang, and Zhiyao Xie. Rtlcoder: Fully open-source and efficient llm-assisted rtl code generation technique. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2024c.

Yao Lu, Shang Liu, Qijun Zhang, and Zhiyao Xie. Rtllm: An open-source benchmark for design rtl generation with large language model. In *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 722–727. IEEE, 2024.

Madhav Nair, Rajat Sadhukhan, and Debdeep Mukhopadhyay. Generating secure hardware using chatgpt resistant to cwes. *Cryptology ePrint Archive*, 2023.

OpenAI. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: September 24, 2025.

OpenAI. Introducing gpt-5, 2025. URL https://openai.com/index/introducing-gpt-5/. Accessed: September 24, 2025.

Zehua Pei, Hui-Ling Zhen, Mingxuan Yuan, Yu Huang, and Bei Yu. Betterv: Controlled verilog generation with discriminative guidance. *arXiv preprint arXiv:2402.03375*, 2024.

Shailja Thakur, Baleegh Ahmad, Zhenxing Fan, Hammond Pearce, Benjamin Tan, Ramesh Karri, Brendan Dolan-Gavitt, and Siddharth Garg. Benchmarking large language models for automated verilog rtl code generation. In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6. IEEE, 2023.

Shailja Thakur, Baleegh Ahmad, Hammond Pearce, Benjamin Tan, Brendan Dolan-Gavitt, Ramesh Karri, and Siddharth Garg. Verigen: A large language model for verilog code generation. *ACM Transactions on Design Automation of Electronic Systems*, 29(3):1–31, 2024.

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

Xi Wang, Gwok-Waa Wan, Sam-Zaak Wong, Layton Zhang, Tianyang Liu, Qi Tian, and Jianmin Ye. Chatcpu: An agile cpu design and verification platform with llm. In *Proceedings of the 61st ACM/IEEE Design Automation Conference*, pp. 1–6, 2024.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhongzhi Yu, Mingjie Liu, Michael Zimmer, Yingyan Celine, Yong Liu, and Haoxing Ren. Spec2rtl-agent: Automated hardware code generation from complex specifications using llm agent systems. In *2025 IEEE International Conference on LLM-Aided Design (ICLAD)*, pp. 37–43. IEEE, 2025.

Yang Zhao, Di Huang, Chongxiao Li, Pengwei Jin, Muxin Song, Yinan Xu, Ziyuan Nan, Mingju Gao, Tianyun Ma, Lei Qi, et al. Codev: Empowering llms with hdl generation through multi-level summarization. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2025.

Yujie Zhao, Hejia Zhang, Hanxian Huang, Zhongming Yu, and Jishen Zhao. Mage: A multi-agent engine for automated rtl code generation. *arXiv preprint arXiv:2412.07822*, 2024.

Yaoyu Zhu, Di Huang, Hanqi Lyu, Xiaoyun Zhang, Chongxiao Li, Wenxuan Shi, Yutong Wu, Jianan Mu, Jinghua Wang, Yang Zhao, et al. Codev-r1: Reasoning-enhanced verilog generation. *arXiv preprint arXiv:2505.24183*, 2025.

A THE SYSTEM LEVEL RESULT OF REALBENCH

Figure 6 presents the design hierarchy of RealBench and the corresponding performance of LocalV. Specifically, it details the verification outcomes for (a) an SD card controller, (b) an AES encoder/decoder core, and (c) the Hummingbirdv2 E203 CPU Core. A "Pass" denotes successful module generation by LocalV, whereas a "Fail" indicates an unsuccessful attempt. The hierarchical tree structure within the figure visually represents the intricate task interdependencies in RealBench, underscoring its inherent complexity.

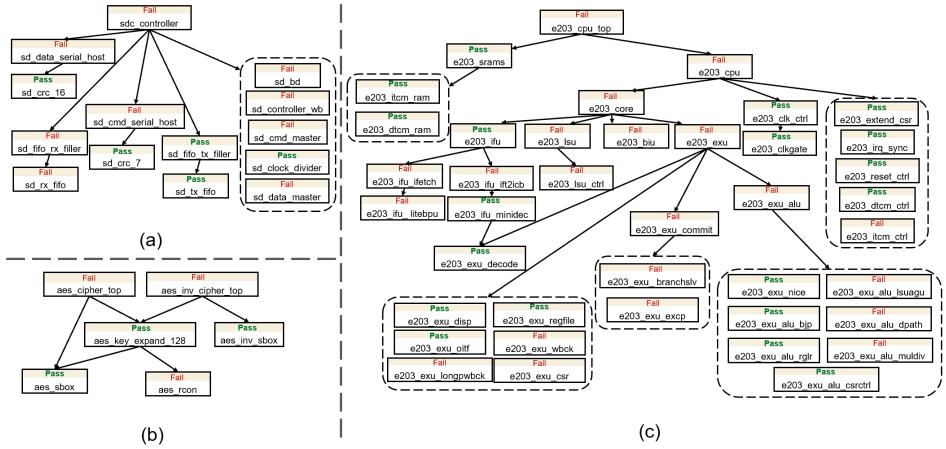


Figure 6: The system level result of RealBench.

B INTERMEDIATE RESULTS OF LOCALV

To better illustrate LocalV's workflow, this section delves into the detailed intermediate results for the **e203_exu** problem in REALBENCH. We'll display the outputs generated by LocalV agents, including document fragments, pseudocode, plans, code fragments, and debug actions, providing a comprehensive understanding of the process.

Document Fragments

"section_0": "The 'e203_exu' module represents the execution unit (EXU) of a processor core, responsible for executing instructions received from the instruction fetch unit (IFU) and handling exceptions, pipeline flushes, and write-back operations. This module integrates submodules for decoding, dispatching, ALU operations, commit handling, and more, ensuring proper execution of instructions and system state management.",
"section_1": "| Direction | Port Name | Width | Description |
———	—————	—————	———————————————————
output	commit_mret	1	Indicates that an MRET instruction has been committed.
output	commit_trap	1	Indicates that a trap (e.g., exception) has been committed.
output	exu_active	1	Indicates whether the execution unit is currently active.
output	excp_active	1	Indicates whether an exception is currently active.
output	core_wfi	1	Indicates that the core is in a waiting-for-interrupt (WFI) state.
output	tm_stop	1	Indicates that the timer should stop.
output	itcm_nohold	1	Indicates no hold on the instruction TCM.
output	core_cgstop	1	Core clock gating stop signal.
output	tcm_cgstop	1	TCM clock gating stop signal.
input	core_mhartid	E203_HART_ID_W	Hardware thread ID of the core.
input	dbg_irq_r	1	Debug interrupt request signal.
....",
.....

Pseudocode

```
// === SUBTASK 1: Module Header and IO Port Definitions ===
module e203_exu (
// <basic interface signals>
// <debug control interface signals>
// <IFU IR stage interface signals>
// <flush interface signals>
// <LSU write-back interface signals>
// <AGU ICB interface signals>
// <optional CSR_NICE interface signals>
// <optional NICE interface signals>
);
// === SUBTASK 2: Internal Signal Declarations ===
// <register file interface signals>
// <decode stage signals>
// <dispatch stage signals>
// <OITF signals>
// <ALU signals>
// <commit stage signals>
// <CSR signals>
// <long-pipeline write-back signals>
// <write-back signals>
// === SUBTASK 3: Submodule Instantiations ===
// Register File
e203_exu_regfile u_e203_exu_regfile (
// <register file connections>
);
// Decode
e203_exu_decode u_e203_exu_decode (
// <decode connections>
);
// Dispatch
e203_exu_disp u_e203_exu_disp (
// <dispatch connections>
);
// OITF
e203_exu_oitf u_e203_exu_oitf (
// <OITF connections>
);
// ALU
e203_exu_alu u_e203_exu_alu (
// <ALU connections>
);
// Commit
e203_exu_commit u_e203_exu_commit (
// <commit connections>
);
// CSR
e203_exu_csr u_e203_exu_csr (
// <CSR connections>
);
// Long-pipeline Write-back
e203_exu_longpwbck u_e203_exu_longpwbck (
// <long-pipeline write-back connections>
);
// Write-back
e203_exu_wbck u_e203_exu_wbck (
// <write-back connections>
);
// === SUBTASK 4: Conditional Logic and Signal Assignments ===
// <exu_active signal logic>
// <dependency information logic>
// <optional NICE interface logic>
// <optional CSR_NICE interface logic>
endmodule
```

Plan

”subproblems”: [
{ ”id”: ”module_header_and_io”, ”description”: ”**Pseudocode Block**: This subtask implements the [SUBTASK 1: Module Header and IO Port Definitions] block in the provided pseudocode. **Target Implementation**: Define the complete module header with all interface signals including basic interface, debug control interface, IFU IR stage interface, flush interface, LSU write-back interface, AGU ICB interface, and optional CSR_NICE and NICE interfaces according to the specifications.”, ”required_section_indexes”: [”section_0”, ”section_1”, ”section_2”, ”section_3”, ”section_4”, ”section_5”, ”section_6”, ”section_7”, ”section_8”] },
{ ”id”: ”internal_signals”, ”description”: ”**Pseudocode Block**: This subtask implements the [SUBTASK 2: Internal Signal Declarations] block in the provided pseudocode. **Target Implementation**: Define all internal signals needed for connecting the submodules, including register file interface signals, decode stage signals, dispatch stage signals, OITF signals, ALU signals, commit stage signals, CSR signals, long-pipeline write-back signals, and write-back signals.”, ”required_section_indexes”: [”section_9”, ”section_10”, ”section_11”, ”section_12”, ”section_13”, ”section_14”, ”section_15”, ”section_16”, ”section_17”] },
.....]

Code Fragment

```
// === SUBTASK 4: Conditional Logic and Signal Assignments ===
// exu_active signal logic
// Active when OITF is not empty, valid instruction entering, or exception active
assign exu_active = (~ oitf_empty) | i_valid | excp_active;
// dependency information logic for IFU
// dec2ifu_rden: Indicates whether current instruction writes to an integer register
assign dec2ifu_rden = dec_rdwen & (~dec_rdwen_fp);
// dec2ifu_rs1en: Indicates whether current instruction reads rs1 integer register
assign dec2ifu_rs1en = dec_rs1en & (~dec_rs1en_fp);
// dec2ifu_rdidx: Destination register index
assign dec2ifu_rdidx = dec_rdidx;
// rf2ifu_rs1: Value of rs1 register
assign rf2ifu_rs1 = read_src1_dat;
`ifdef E203_HAS_NICE
// optional NICE interface logic
assign nice_cmt_off_ilgl_o = nice_cmt_off_ilgl_i;
assign nice_xs_off = nice_xs_off_i;
`endif
`ifdef E203_HAS_CSR_NICE
// optional CSR_NICE interface logic
assign csr_nice_csr_valid = csr_nice_csr_valid_i;
assign csr_nice_csr_ready = csr_nice_csr_ready_i;
assign csr_nice_csr_addr = csr_nice_csr_addr_i;
assign csr_nice_csr_wr = csr_nice_csr_wr_i;
assign csr_nice_csr_wdata = csr_nice_csr_wdata_i;
assign csr_nice_csr_rdata = csr_nice_csr_rdata_i;
`endif
```

Debug Action

”fix_operations”: [{ ”operation”: ”delete_block”, ”start_line”: 838, ”end_line”: 843 }, { ”operation”: ”add_block”, ”line”: 838, ”content”: [” // Flush interface logic - handle different flush conditions”, ” // Flush can be triggered by exceptions, interrupts, or control flow instructions”, ” assign pipe_flush_req = commit_trap | commit_mret;”, ” ”, ” // For normal exceptions and interrupts, use the trap vector”, ” // For returns from exceptions (mret), use the saved EPC”, ” assign pipe_flush_add_op1 = commit_mret ? csr_epc_r : csr_mtvec_r;”, ” assign pipe_flush_add_op2 = 32’b0;”, ”‘ifdef E203_TIMING_BOOST”, ” assign pipe_flush_pc = commit_mret ? csr_epc_r : csr_mtvec_r;”, ”‘endif”] }, { ”operation”: ”delete_block”, ”start_line”: 817, ”end_line”: 819 }, { ”operation”: ”add_block”, ”line”: 817, ”content”: [” // exu_active signal logic”, ” // Active when OITF is not empty, valid instruction entering, or exception active”, ” // Also active during flush operations”, ” assign exu_active = (~oitf_empty) | i_valid | excp_active | pipe_flush_req;”] }],

C LLM USAGE

Large language models (LLMs) were utilized to assist in the writing and polishing of this manuscript. Specifically, LLMs were employed to help refine language, improve readability, and enhance clarity across various sections of the paper. This included tasks such as rephrasing sentences, checking grammar, and improving the overall coherence and flow of the text.