Poster Contacts: yundat@nvidia.com, mingjiel@nvidia.com, haoxingr@nvidia.com

NVIDIA

CraftRTL: High-quality Synthetic Data Generation for Verilog Code Models with Correct-by-Construction Non-Textual Representations and Targeted Code Repair

Mingjie Liu* · Yun-Da Tsai* · Wenfei Zhou · Haoxing Ren

We present **CraftRTL**, a state-of-the-art Hardware LLM designed for RTL coding. Our approach enhance data curation by generating correct-by-construction data targeting non-textual representations. Additionally, we introduce an automated framework that gathers errors from multiple model checkpoints and create targeted code repair data by error injection. We outperform prior results by **3.8%**, **10.9%**, **6.6%** for pass@1 on VerilogEval and RTLLM.

Motivation

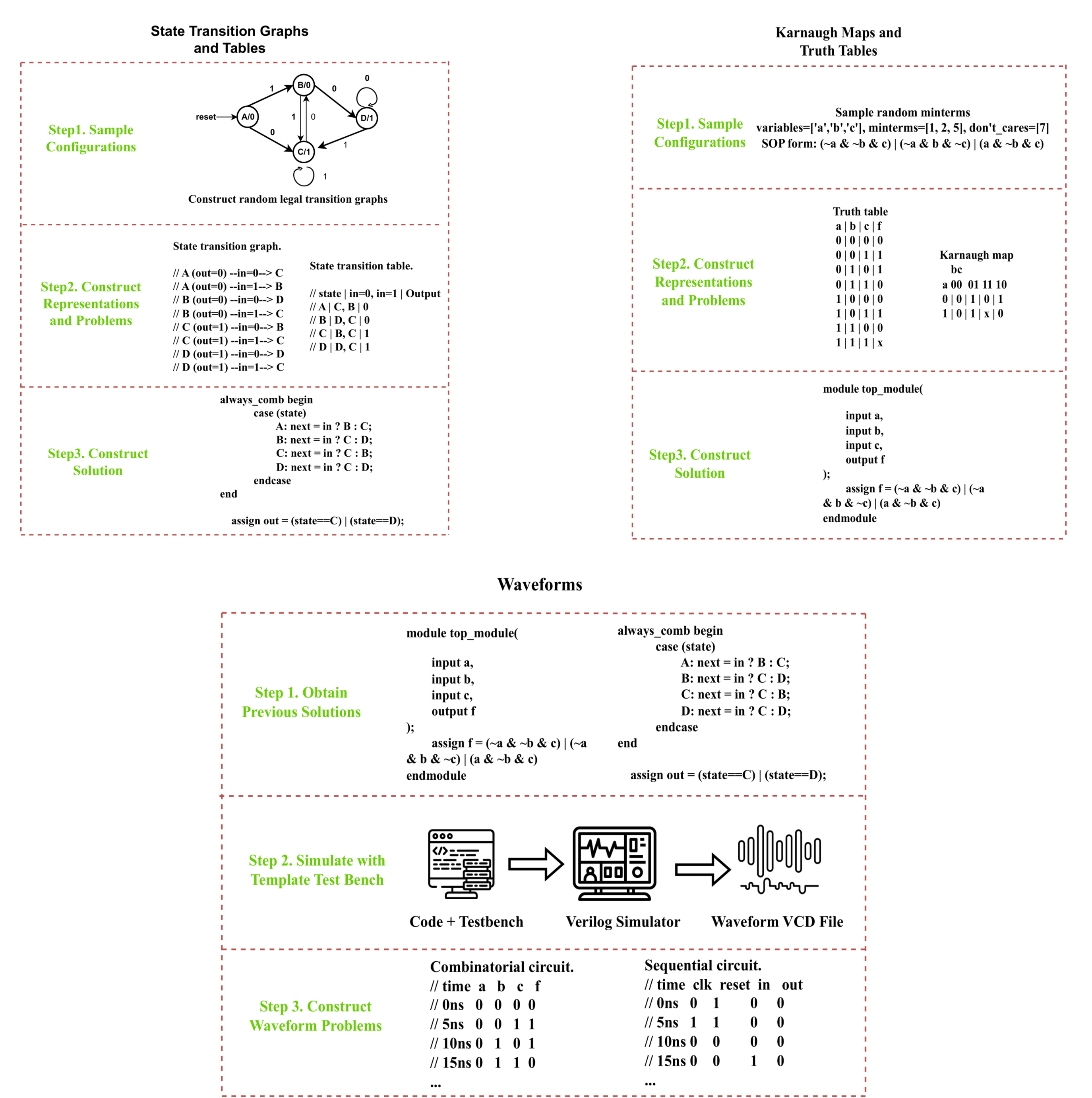
- **Challenges:** Limited high-quality data, constrained capability of existing LLMs and complex verification requirements for ensuring the functional correctness.
- **LLMs for RTL Generation:** Reducing designer costs and efforts for completing hardware specs, code synthesis, refinement and debugging.
- **Prior works on Synthetic Data Generation:**
  1) Self-Instruct generate with in-context learning from seed examples (Wang, Yizhong, et al. 2022)
  2) Docu-Instruct utilize document sources from Wikipedia and textbooks to enhance diversity (Adler, Bo, et al. 2024)
  3) Oss-Instruct reversely generate coding problems from code snippets (Wei, Yuxiang, et al. 2023)

  Struggle to generate functionally correct data

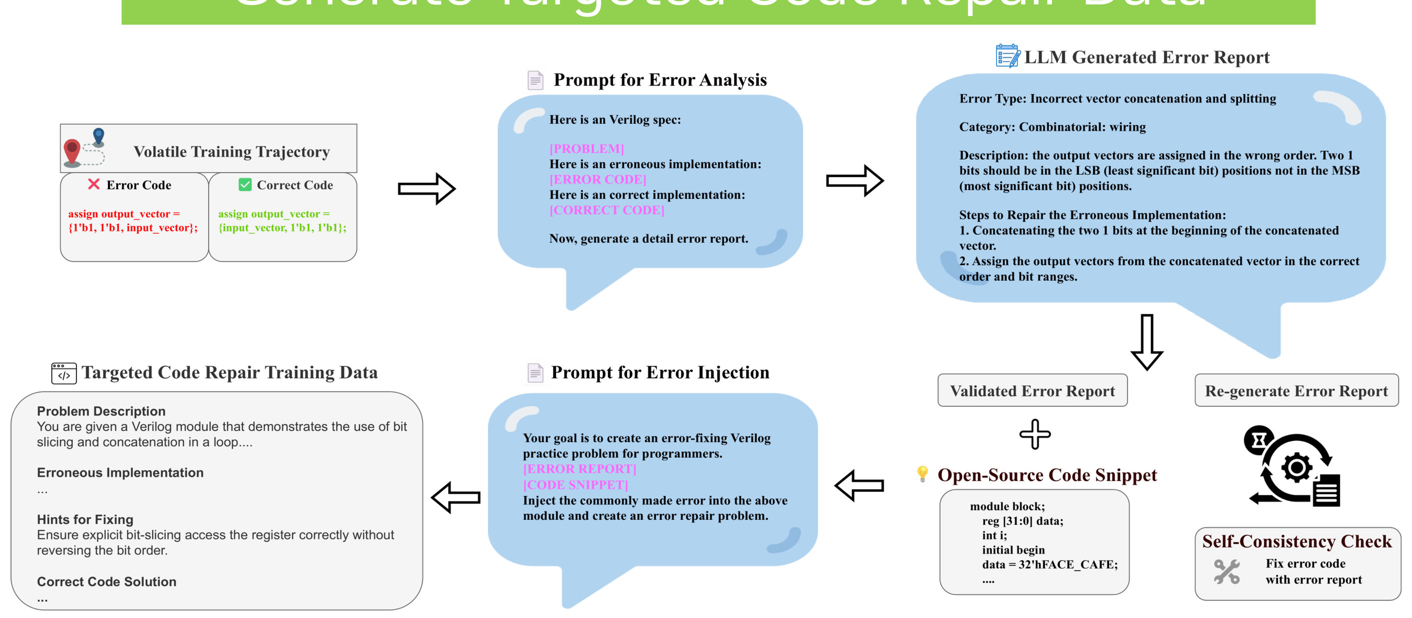
- **Our Contributions:**
  - We develop correct-by-construction datasets targeting non-textual representations, improving model understanding and generation capabilities.
  - We have developed an automated system that collects errors from various model checkpoints and injects these into code to create targeted code repair datasets..
  - CraftRTL-15B surpasses prior state-of-the-art results by achieving higher pass@1 rates on VerilogEval-Machine, VerilogEval-Human, and RTLLM benchmarks.

CraftRTL

Ensure Correctness with Correct-by-Construction



Generate Targeted Code Repair Data



Experimental Results

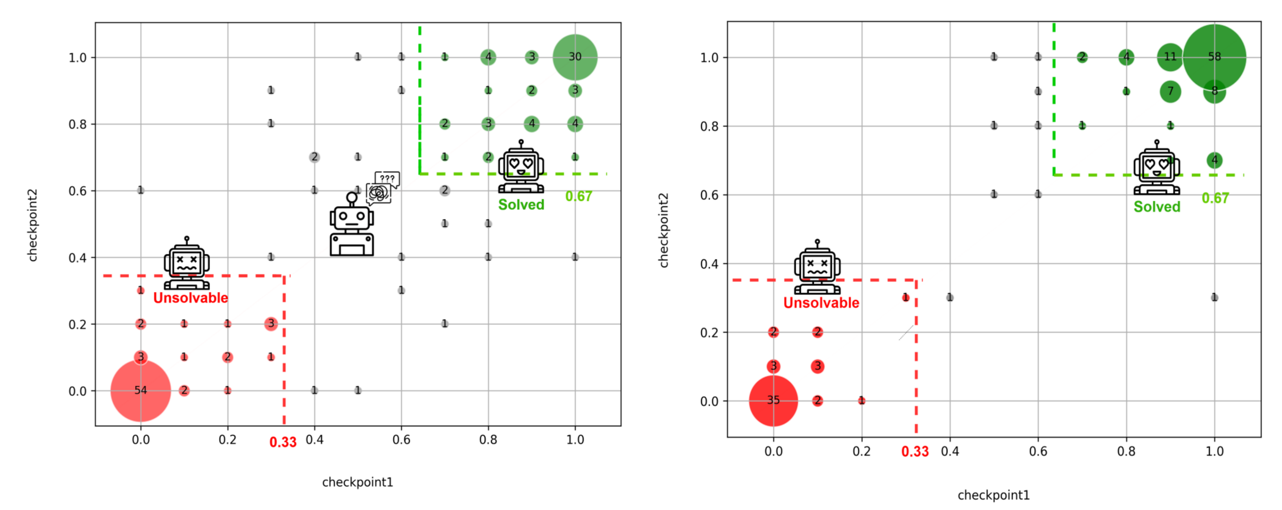
Type	Model	Size	VerilogEval (Liu et al., 2023b) Machine (%) pass@1	pass@5	pass@10	Human (%) pass@1	pass@5	pass@10
Foundational Models	Llama-3.1	8B	48.7	67.3	74.1	26.9	37.8	44.2
	Llama-3.1	405B	67.3	75.1	76.9	53.8	61.0	62.8
	Nemotron-4	340B	53.0	60.3	62.2	43.1	48.3	50.0
	GPT-3.5-turbo	-	58.0	74.0	77.6	31.2	44.1	47.4
	GPT-4o	-	65.9	71.4	72.7	57.1	63.9	66.7
Code Models	CodeLlama	7B	43.1	47.1	47.7	18.2	22.7	24.3
	CodeQwen	7B	46.5	54.9	56.4	22.5	26.1	28.0
	Starcoder2	15B	68.7	82.3	88.5	37.7	50.6	57.2
	DeepSeek-Coder	6.7B	52.2	55.4	56.8	30.2	33.9	34.9
	DeepSeek-Coder-V2	16B	67.4	78.3	81.8	46.9	55.9	58.9
	DeepSeek-Coder-V2	236B	68.2	74.1	76.2	56.4	62.2	66.0
RTLCoder (Liu et al., 2023c)	Mistral	7B	62.5	72.2	76.6	36.7	45.5	49.2
	DeepSeek-Coder	7B	61.2	76.5	81.8	41.6	50.1	53.4
BetterV (Pei et al., 2024)	CodeLlama	7B	64.2	75.4	79.1	40.9	50.0	53.3
	DeepSeek-Coder	6.7B	67.8	79.1	84.0	45.9	53.3	57.6
	CodeQwen	7B	68.1	79.4	84.5	46.1	53.7	58.2
CodeV (Zhao et al., 2024)	CodeLlama	7B	78.1	86.0	88.5	45.2	59.5	63.8
	DeepSeek-Coder	6.7B	77.9	**88.6**	**90.7**	52.7	62.5	67.3
	CodeQwen	7B	77.6	88.2	**90.7**	53.2	65.1	68.5
OriGen (Cui et al., 2024)	DeepSeek-Coder	6.7B	74.1	82.4	85.7	54.4	60.1	64.2
Ours SDG-CC-Repair	CodeLlama	7B	78.1	85.5	87.8	63.1	67.8	69.7
	DeepSeek-Coder	6.7B	77.8	85.5	88.1	65.4	70.0	72.1
	Starcoder2	15B	**81.9**	86.9	88.1	**68.0**	**72.4**	**74.6**

Main Experiment: Pass-rates of recent large language models and the proposed CraftRTL. Pass@1 = #passed case/#total case. We report the best pass@k score among temperature {0.2, 0.8}.

Model	VerilogEval Machine pass@1 (%)	VerilogEval Human pass@1 (%)	RTLLM v1.1 Func pass@5 (%)
Starcoder2-15B	68.7	37.7	37.6
SDG (80.1k)	75.2	54.7	62.1
SDG-CC (108.6k)	73.9	62.0	62.8
SDG-CC-Repair (110.0k)	**81.9**	**68.0**	**65.8**

Model	VerilogEval Machine pass@1 (%)	VerilogEval Human pass@1 (%)	RTLLM v1.1 Func pass@5 (%)
SDG-CC	73.9	62.0	62.8
SDG-CC-Repair	**81.9**	**68.0**	**65.8**
w/o self-consistency	75.3	63.3	63.7
w/o error report	76.9	59.6	59.4

Ablation Study: Pass-rate (%) from the ablation study evaluating the effectiveness of various synthetic data groups and framework components.



Improved Pass-rate variability