

# ChipVQA: Benchmarking Visual Language Models for Chip Design

Haoyu Yang, Qijing Huang, Nathaniel Pinckney, Walker Turner, Wenfei Zhou,  
Yanqing Zhang, Chia-Tung Ho, Chen-Chia Chang, Haoxing Ren  
NVIDIA Corp.

{haoyuy, haoxingr}@nvidia.com

**Abstract**—Large-language models (LLMs) have exhibited great potential to assist chip designs and analysis. Recent research and efforts are mainly focusing on text-based tasks including general QA, debugging, design tool scripting, and so on. However, chip design and implementation workflow usually require a visual understanding of diagrams, flow charts, graphs, schematics, waveforms, etc, which demands the development of multi-modality foundation models. In this paper, we propose ChipVQA, a benchmark designed to evaluate the capability of visual language models for chip design. ChipVQA includes 142 carefully designed and collected VQA questions covering five chip design disciplines: *Digital Design*, *Analog Design*, *Architecture*, *Physical Design* and *Semiconductor Manufacturing*. Unlike existing VQA benchmarks, ChipVQA questions are carefully designed by chip design experts and require in-depth domain knowledge and reasoning to solve. We conduct comprehensive evaluations on both open-source and proprietary multi-modal models that are greatly challenged by the benchmark suit. ChipVQA is available at <https://github.com/phdyang007/chipvqa>.

## I. INTRODUCTION

In the past decade, AI for chip design has gathered significant focus, revolutionizing the semiconductor industry. In particular, supervised learning techniques are employed to identify and mitigate hotspots [1], defects [2], [3] in chip layouts, ensuring higher reliability and performance. On the other hand, generative models are used for data generation and optimization [4], [5], automating the creation of innovative chip architectures and improving design efficiency.

Exploding of large language models (LLMs) have brought new potential to assist chip design and manufacturing flows. Thanks to their capabilities in producing human-like natural language, logical reasoning, and coding, LLMs have been investigated on application scenarios of engineering assistant chatbot, EDA script generation, bug analysis [6], RTL code completion [7], [8] and bug fix [9]. Though there are continuing progresses on LLM-Aided chip design, we are facing two major challenges: 1) **Lacking support of multi-modality**. Existing LLM development for chip designs are purely text-based framework while multiple modalities drastically occur in entire chip design cycle: diagrams, graph, schematic, layouts, waveform, table, curve and so on. Capability of understanding modalities other than text will significantly boost the LLM application scenario in chip design. 2) **Lacking benchmark**. Benchmark suites play an important role on machine learning model development. A high quality benchmark collection offers standard evaluation and guidance for model development.

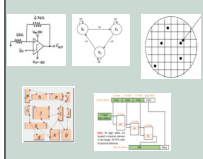
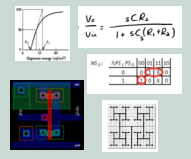
Broad Discipline	Diverse Visual Content	Comprehensive Difficulty
Analog, Digital, Architecture, Physical Design, Manufacture	Schematic, Graph, Equation, Table, Waveform, Flow Chart,...	Undergraduate, Graduate, Academic Research, Industry
		<ul style="list-style-type: none"> <li>Logic Efforts</li> <li>Yield Analysis</li> <li>Congestion Map</li> <li>Small-Signal Analysis</li> <li>Karnaugh Map</li> <li>Doping, Etching</li> <li>Placement, Routing</li> <li>Pipeline, Cache</li> </ul>

Fig. 1 ChipVQA features broad knowledge disciplines, diverse visual content and comprehensive difficulties for VLM benchmark.

Though there are benchmark efforts related to hardware design like VerilogEval [10], the knowledge scope is quite limited to RTL design. Particularly, there is no descent benchmark suite focuses on broad chip design knowledge base with multi-modality (visual) contents.

In this paper, we introduce ChipVQA (as shown in Fig. 1), a comprehensive VQA benchmark suit covering broad chip design disciplines: *Digital Design*, *Analog Design*, *Architecture*, *Physical Design* and *Semiconductor Manufacturing*. ChipVQA comprises 142 meticulously crafted VQA questions, developed by seasoned chip design experts, each with over ten years of industry experience. This ensures the benchmark's high quality, as the questions span a wide range of topics and challenge various aspects of chip design knowledge, reasoning, and problem-solving skills. The comprehensive nature of these questions provides a robust and reliable assessment tool for evaluating the capabilities of visual-language models in the context of chip design. Specifically, unlike existing benchmark efforts targeting at most undergraduate level engineering question [11], ChipVQA questions encompass a range from college courses to practical research topics, demanding extensive knowledge in chip design, as well as deduction and reasoning skills to answer effectively. ChipVQA also includes a diverse array of visual content, such as diagrams, schematics, equations, pictures, flow charts, tables, layouts, and more. This variety poses a substantial challenge to state-of-the-art visual-language models like GPT-4o [12].

We evaluated ten popular open source VLM models and the state-of-the-art proprietary model GPT-4o on ChipVQA and discovered:

- ChipVQA challenges existing VLMs on chip design

domain knowledge understanding and reasoning. In particular, GPT-4o achieves only 44% correctness rate.

- For chip visual question answering, the correctness rate of visual-language models (VLMs) is largely determined by their text processing capabilities (the LLM component). This is demonstrated by our experiments on LLaVA [13], [14] with different language model backbones.
- We examined ChipVQA using various image resolutions and concluded that higher resolution images improve the effectiveness of visual question answering, as exemplified in MM1 [15].
- All VLMs exhibit higher performance on multiple choice questions than short answers, when the answer options in the prompt to some extent offer retrieval augmented generation (RAG).
- Lastly, we conducted a preliminary study on an alternative VLM inference methodology using an agent [16], which showed improved performance without additional training on certain categories. This highlights the potential of utilizing LLM agents as an alternative approach for VLM deployment in chip design.

The reminders of the paper are organized as follows: Section II introduces related works; Section III details the benchmark data collection and configurations; Section IV presents the experimental results of our benchmark on state-of-the-art visual-language models; and Section V concludes the paper and future work.

## II. RELATED WORKS

### A. Visual Language Models

VLMs are designed to understand and generate natural language based on visual inputs and text prompts [12]–[15], [17]–[21], which are expected to handle various visual tasks such as captioning, visual question answering (VQA), image generation and document comprehension. Capable of dealing with visual content brings VLM closer towards next level artificial general intelligence (AGI). A representative VLM pipeline is depicted in Fig. 2, which has three major components:

- 1) A visual encoder that performs image processing and extract visual embedding,
- 2) A projection unit that convert visual embedding into tokens that are compatible with text tokens handled by large language model, and
- 3) A large language model takes input of both visual and text tokens and generate output texts.

Training and deploying Vision-Language Models (VLMs) requires substantial effort and investment, highlighting the difficulty and expense involved in establishing a foundational data pipeline as in Fig. 2. In particular, these efforts include preparation of visual-text alignment data and the demands of computing resources for VLM training [15], given that the capabilities of VLMs are largely determined by the capacity of the underlying LLM. There are also researches that enable multi-modality output with decoder attached (mostly diffusion models) following LLM outputs e.g. [12], [14], [22].

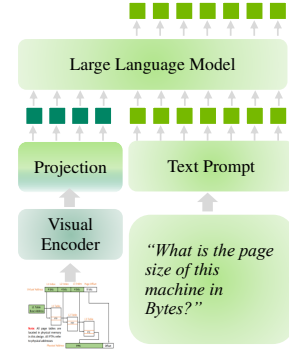


Fig. 2 Representative architecture of visual-language-models.

### B. Engineering VQA Benchmarks

The rapid growth of Vision-Language Models (VLMs) has spurred investigations into their capacity to address STEM problems, necessitating the development of comprehensive benchmarks. Notable examples include MMBench [23], MM-Vet [24], LAMM [25], SEED [26], MathVista [27], and MMMU [11], which encompass a wide array of engineering and scientific content. However, the majority of these benchmark questions are constrained by domain knowledge depth, primarily covering content up to the level of undergraduate engineering courses. Specifically, existing benchmarks do not adequately cover the comprehensive domain content related to chip design, implementation, and manufacturing Fig. 3. Consequently, they fall short in evaluating the capability of VLMs to assist with practical applications in the chip and semiconductor industry. ChipVQA differs from prior arts with VQA collection spanning key chip design disciplines that cover difficulty level from undergraduate course to practical industry contents. More specifically, all questions are manually collected by domain experts with 10+ years of experience, ensuring the quality and depth of the benchmark. Notably, GPT-4o only achieves 44% pass@1 on ChipVQA standard collection and 20% pass@1 on ChipVQA challenge collection (all multiple-choice questions replaced with short answer questions), as compared to 69.1% pass@1 on MMMU [11].

## III. CHIPVQA BENCHMARK

### A. Benchmark Overview

Given the significant potential of VLMs in aiding chip design and the current lack of comprehensive evaluation, we introduce ChipVQA, a curated set of manually crafted visual-question-answer triplets designed for benchmarking VLMs. ChipVQA distinguishes itself from existing engineering benchmarks, as outlined in Fig. 1. Specifically, our focus spans across various facets of chip design including digital design, analog design, architecture, physical design, and chip manufacturing, with each area curated by multiple domain experts.

1) *Statistics*: Key statistics for ChipVQA are detailed in TABLE I, comprising 142 meticulously curated chip design questions. Each question is paired with at least one visual component essential for deriving the answer. The benchmark

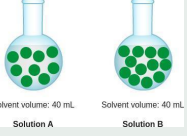
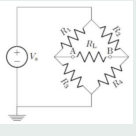
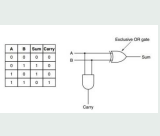
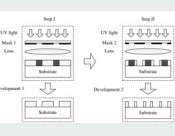
Benchmarks	MMBench	MM-Vet	MathVista	MMMU	ChipVQA
Question	Which solution has a higher concentration of green particles? HINT: The diagram below is a model of two solutions. Each green ball represents one particle of solute.	What is $d$ in the last equation?	Given $V_s = 5V$ , $R_1 = 1k\Omega$ , $R_2 = 2.2k\Omega$ , $R_3 = 2.2k\Omega$ , $R_4 = 1.5k\Omega$ , and $R_L = 4.7k\Omega$ . Determine the voltage and current across $R_L$ . Answer in unit of V.	The <image 1> shows the truth table and calculation circuit diagram for the addition of 1-digit integers. What is the simple circuit that the diagram represents usually called?	What is the lithography resolution enhancement technique depicted in the figure?
Visual		Solve the following equations: 1) $8x + 11 = 4x + 14$ 2) $7d - 4 = 11d - 9$			

Fig. 3 Question samples from existing engineering VQA benchmarks and ChipVQA.

includes two question types: multiple choice, where each question is accompanied by four answer options presented as text prompts; and short answer, requiring open-ended responses such as numerical values or brief explanations. Our collection features 12 distinct types of visual content, ensuring a thorough evaluation. Among these, "schematic", "diagram", and "layout" constitute the majority, reflecting their frequent presence in chip design materials and documents. The prompts in each question are crafted with tokens of varying lengths, from 5 to 370 tokens, presenting challenges of different complexities in context comprehension.

TABLE I Statistics of ChipVQA

Data	Total	MC	SA
	142	99	43
Category	Type	Count	
	Digital Design	35	
	Analog Design	44	
	Architecture	20	
	Manufacture	20	
	Physical Design	23	
Visual	Type	Count	
	schematic	53	
	diagram	29	
	layout	16	
	table	5	
	mixed	5	
	structure	3	
	figure	3	
	curve	4	
	flow	1	
	equations	1	
	neural nets	1	
	equation	1	
Prompt Token	Item	Token Length	
	mean	33.48	
	std	54.87	
	min	5	
	25%	11	
	50%	16	
	75%	28.75	
	max	370	

2) *Quality and Copyright*: To ensure a high-quality benchmark, all VQA samples are collected and designed from professional materials including textbooks, research manuscripts, course exams, and open course materials. A large fraction of questions and answers are (re)designed by domain experts (including co-authors) that grant the desired level of difficulty, which costs over 200 human-hour for question collection and annotation. We carefully reviewed all referenced materials to comply with associated copyrights and regulations. Specifically, we acquired written approval from authors/publishers on

using proprietary content and filtered out content prohibited for distribution.

## B. Data Collection

1) *Digital Design*: The Digital Design multiple-choice question dataset comprises 35 triplets, each consisting of a prompt, a figure, and four answer choices. These questions cover a broad range of introductory-level digital design topics typically encountered in the first two years of undergraduate computer science or electrical engineering curricula, with some advanced questions at the graduate level. The topics include Processor and CPU Design, Logic Design and Circuit Analysis, Data Representation, Functional Derivation, and Memory and Storage Design.

Expert annotators, defined as individuals with at least graduate-level training in Computer Science or Electrical Engineering and Computer Science (CS/EECS), collected data from open-source textbooks and exam papers. The annotators constructed the multiple-choice questions using the following methods:

- Direct extraction from the source material
- Manual creation of question and multiple choice answer pairs based on the source materials
- Masking parts of the image or textbook description to create alternative plausible solutions

To ensure data quality, the questions and answers were further validated by seasoned chip designers with doctoral degrees and more than 10 years of experience in the field. Special attention was given to ensure that the answer choices are syntactically and even semantically similar to each other, as well as logically plausible. For example, the question "Derive the function for  $Q$  given the state table and excitation maps as shown in the figures" has four answer choices, all of which could be inferred, but only one is correct: a)  $Q = S'Q + S$ , b)  $Q = S'R'q + SR'$ , c)  $Q = SR' + R'q$ , d)  $Q = S'Q + SR'$ . In so doing, the approach increases challenges the LLM to truly understand what is being asked and gives well-thought-out answers to the question.

2) *Analog Design*: 44 multiple-choice questions were used for the Analog Design dataset. The questions span 38 unique images comprising various amplifier-level and transistor-level schematics as well as Bode plots and symbolic trans-

fer functions. The circuit schematic images include amplifiers, negative feedback networks, analog-to-digital converters (FLASH, SAR, pipeline residue amplifiers), comparator-based oscillators, and instrumentation amplifiers. Question topics cover DC operating points, small-signal gain, equivalent resistance, closed-loop feedback analysis, transfer functions, pole/zero/unity-gain frequencies, phase margin, circuit voltage range, and compensation. Additional context is provided within some images such as device parameters, transfer functions, and device labels.

3) *Architecture*: The computer architecture dataset consists of 21 short-answer and multiple-choice questions. The questions and answers are paired with relevant images sourced from exams and exercises online. The image selection aimed for broad coverage, encompassing topics including memory encoding, branch prediction, critical path latency, coherence protocol, virtual memory translation, pipelining, vector processor, out-of-order machines, network topology, etc. The dataset was created through manual collection. The authors meticulously reviewed each exam and exercise question, capturing figures that represented the subject matter. The authors self-assessed the difficulty of the manually created QA points as appropriate for graduate-level computer architecture students. The images serve as crucial visual representations of the problems, aiding the model in comprehending the questions and formulating accurate answers. For example, a question might ask how a bolded bypass path in a pipeline diagram (connecting the load unit output to the ALU input) affects the cycle per instruction and frequency.

4) *Physical Design*: Short answer and multiple choice questions and answers comprise the dataset for the Physical Design section. Images were gathered from textbooks that cover a wide range of physical design topics. The method for dataset generation was manual collection, by the authors, by going through each textbook and capturing figures where the authors immediately understood the subject matter. The authors then manually created a QA point based on their own understanding of the subject, and a quick review based on the context of the figure at hand. The manually created QA difficulty was self-assessed by the authors to be of graduate school, VLSI-related major level. Emphasis was placed on covering a wide range of physical design topics, including clock tree generation, clock tree and signal routing, routing algorithms, physical design methodology and flow overview, standard cell layout and pin access, DRC, power/ground network design, standard cell placement and legalization, floor planning and macro placement, timing analysis and useful skew, and logic restructuring.

In this way, a total of 23 questions are collected. The images provide the model additional context to reason through, while the question text sets the topic space for the model to aid in its inference. For example, a question may be "The routing points' coordinates are shown, can you calculate the routing costs for the 2 diagrams and determine which routing topology has lower cost?", while the accompanying image provides a

Steiner tree diagram with annotated routing points.

5) *Chip Manufacturing*: We collected 20 VQA problems related to chip manufacturing, including both multiple-choice and short-answer questions. These questions were sourced from open materials, textbooks, and research documents. The topics cover a wide range of semiconductor manufacturing aspects, such as lithography, solid-state physics, ion deposition, wafer defects, etching, doping, device layouts, and more. To ensure quality and appropriate difficulty levels, most questions were produced by domain experts through manual annotation. We meticulously reviewed and refined the textual and visual content from the materials to generate questions that require domain knowledge, reasoning, and arithmetic deduction. For example, consider the following question prompt: "Assume 5:1 BOE (Buffered HF) etches  $\text{SiO}_2$  isotropically at 100 nm/min, RIE etches  $\text{SiO}_2$  at 200 nm/min and has a  $\text{SiO}_2$ :Si selectivity of 15:1, Assume a Si/SiO<sub>2</sub> substrate with patterned photoresist as shown in the figure. For the structure above, how long should this wafer be placed in 5:1 BOE etchant to record a 10% over-etch?" accompanied by an image of a silicon structure, a VLM must understand the basic etching process, recognize the silicon structure, and perform arithmetic computations to derive the final answer.

#### IV. EXPERIMENTAL RESULTS

We evaluate multiple open-source and proprietary VLMs [12]–[14], [17]–[21], [28], [29] using ChipVQA. Major studies are based on a zero-shot inference scheme where we pick up the latest checkpoints of related VLMs without alignment/instruction fine-tuning. For supported VLMs, we provide a separate system prompt for question-answering. For VLMs that do not support system prompts, e.g. Paligemma [20], the original system prompt will be concatenated with the user question prompt together as instructions. We also examine the VLMs capability by upgrading the challenge that we replaced all multiple-choice questions with short answer questions. A few rounds of prompt engineering are conducted to guarantee that the VLMs will follow instructions in best practice. Our test platform includes local deployment through Ollama [30] (LLaVA Series), NVIDIA NIM [31] (Neva, Fuyu, Paligemma, Kosmos-2, Phi-3-Vision, VILA, and LLaMA-3.2), and Azure OpenAI [32] (GPT-4o). For all VLMs, we choose temperature=0.1 to preserve the deterministic model output.

We incorporate a hybrid evaluation to check the correctness of the model response. First, we have an auto-evaluation flow based on GPT-4, which will be prompted to check the equivalence of the model response and the golden answer. To do so, we provide a system prompt for GPT-4 and instruct it to provide a binary answer as to whether two responses are equivalent. For certain questions that require access to the QA prompt and/or visual contents to be judged, we conduct manual checks by the annotators to ensure an accurate evaluation.

##### A. Zero-Shot Inference on VLMs

In our initial experiment, we evaluate the zero-shot performance of various Vision-Language Models (VLMs), detailed



TABLE II Zero-Shot Evaluation on ChipVQA

Model	w/ Multi-Choice						w/o Multi-Choice					
	Digital	Analog	Architecture	Manufacture	Physical	all	Digital	Analog	Architecture	Manufacture	Physical	all
LLaVA-7b [14]	0.37	0.20	0.20	0.05	0.22	0.22	0.03	0.00	0.10	0.05	0.09	0.04
LLaVA-13b [14]	0.23	0.16	0.25	0.10	0.17	0.18	0.00	0.02	0.20	0.15	0.04	0.06
LLaVA-34b [14]	0.26	0.32	0.20	0.15	0.22	0.24	0.06	0.05	0.10	0.15	0.17	0.09
LLaVA-LLaMa-3 [14]	0.37	0.18	0.30	0.20	0.22	0.25	0.03	0.00	0.15	0.05	0.13	0.06
NeVA-22b [18]	0.37	0.23	0.15	0.05	0.22	0.22	0.03	0.07	0.10	0.20	0.04	0.08
fuyu-8b [21]	0.11	0.30	0.10	0.05	0.13	0.16	0.00	0.00	0.05	0.05	0.13	0.03
paligemma [20]	0.03	0.07	0.15	0.20	0.04	0.08	0.03	0.00	0.05	0.05	0.04	0.03
kosmos-2 [17]	0.06	0.00	0.05	0.05	0.00	0.03	0.03	0.02	0.00	0.05	0.09	0.03
phi3-vision [19]	0.29	0.18	0.10	0.10	0.30	0.20	0.09	0.05	0.00	0.15	0.17	0.08
VILA-Yi-34B [28]	0.43	0.36	0.30	0.05	0.17	0.29	0.06	0.02	0.25	0.00	0.22	0.09
LLaMA-3.2-90B [33]	0.37	0.25	0.15	0.35	0.48	0.31	0.06	0.09	0.10	0.35	0.39	0.09
GPT4o [12]	0.49	0.51	0.30	0.20	0.61	0.44	0.17	0.09	0.15	0.30	0.48	0.20

in TABLE II. The numbers in the table indicate the Pass@1 rate of each VLM across different VQA disciplines. Notably, the LLaVA-NEXT series consistently demonstrates superior performance compared to other open-source VLMs. The only other model that approaches its performance level is phi3-vision. Within the LLaVA case study, we observe that an enhanced LLM backbone generally enhances performance, particularly aligned with the text capabilities across Mistral-7b [34], Vicuna-13b [35], Yi-34b [36], and LLaMa-3-8b [33]. There exists a notable performance gap between proprietary models and their open-source counterparts. Specifically, our results highlight GPT-4o as leading other open-source models by an average of 20% [12].

During the data collection stage, we do not deliberately regulate the proportion of multiple-choice questions versus short-answer questions. Consequently, we notice that certain models exhibit notable performance distinctions across specific categories, as evidenced in the metrics for "Digital" and "Manufacture". The "Digital" category, characterized by a significant prevalence of multiple-choice questions, establishes a baseline pass rate of 25%. In contrast, the "Manufacture" category features a higher concentration of short-answer questions, necessitating more extensive reasoning and deductive capabilities. Additionally, choice candidates in multiple-choice question prompts provide augmented generation that yields higher pass rate. This is similar to the effectiveness of retrieval-augmented-generation (RAG).

To delve deeper into our investigation, we systematically replaced all multiple-choice questions in the ChipVQA collection with short-answer questions. In these revised questions, the prompts remain unchanged, but all answer choices were removed. Models were then directed to generate short answers corresponding to these questions. Under this challenging setting, we obtain a new group of results, where we observe a significant performance drop on all models. Particularly, the average pass rate of GPT-4o drops from 44% to 20%, showing the improvement opportunity for VLMs to understand and assist chip designs.

### B. Study of Visual Content Quality

VQA examines how well a model can perceive the visual contents and integrate them into the reasoning and deduction process. One common question is the importance of the

TABLE III Evaluation of Agent System on ChipVQA.

Collection	Model	Pass@1
With Choice	GPT4o	0.44
	Agent	0.49
No Choice	GPT4o	0.20
	Agent	0.21

image quality. We investigate this by conducting evaluation on ChipVQA samples with different resolution sizes. For simplicity, we use the "Digital" category as a case study and perform the evaluation using GPT-4o, where the original images are down-sampled 8× and 16× respectively. While 8× downsampling still preserves the evaluation pass rate on the original resolution, 16× lower resolution drops the pass rate from 49% to 37%.

### C. Study of VQA Through Agent

Previous research indicates that the ultimate performance of Vision-Language Models (VLMs) is significantly influenced by the text-processing capabilities of Large Language Models (LLMs), which generally depend on the size of the LLM. Consequently, achieving high-performance VLMs often necessitates substantial resource investment during training. Conversely, agent systems, as discussed in [16], are believed to enhance LLM performance without the need for extensive training by leveraging planning and tool-call mechanisms.

For proof of concept, we employ a straightforward setup where a GPT-4-Turbo agent acts as a chip designer without visual access, while GPT-4o functions as a tool to parse and provide visual information content. During the evaluation, the chip designer interprets the text prompts in the question and invokes the tool-call when necessary. The tool (GPT-4o) then describes the visual content. This interactive process repeats until the chip designer arrives at an answer to the question. Additional results are presented in TABLE III, demonstrating a general performance improvement in a significant portion of cases. However, we observed a decrease in pass rates in certain scenarios, particularly in the manufacturing category. A potential explanation is that the chip designer (GPT-4-Turbo) lacks direct access to visual information, relying solely on the descriptive responses provided by the tool (GPT-4o). This observation underscores the need for further investigation into agent-based LLM systems.

## V. CONCLUSION

We introduce `ChipVQA`, a comprehensive VQA benchmark suite designed to evaluate the chip design capabilities of visual-language models (VLMs). Our dataset comprises 142 VQA questions across five key disciplines related to chip design, with all questions and answers annotated by experienced domain experts. We conducted extensive evaluations on various VLMs, including both open-source and proprietary models. Our results reveal that the benchmark poses significant challenges to state-of-the-art VLMs, as evidenced by their notably low pass rates. Additionally, we investigated several factors that influence VLM performance. Furthermore, we present a preliminary study on an agent-based VQA system, which demonstrates promising potential to enhance LLM/VLM problem-solving capabilities with minimal training overhead. Benchmark samples are open-sourced, with the full collection to be made available upon the final paper release. Future works include `ChipVQA`-oriented dataset collection, VLM training and development, targeting a low-cost yet effective open-source foundation model release.

## REFERENCES

- [1] Z. Xie, Y.-H. Huang, G.-Q. Fang, H. Ren, S.-Y. Fang, Y. Chen, and N. Corporation, "RouteNet: Routability prediction for mixed-size designs using convolutional neural network," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2018.
- [2] H. Yang, J. Su, Y. Zou, B. Yu, and E. F. Y. Young, "Layout hotspot detection with feature tensor generation and deep biased learning," in *ACM/IEEE Design Automation Conference (DAC)*, 2017, pp. 62:1–62:6.
- [3] H. Geng, H. Yang, L. Zhang, J. Miao, F. Yang, X. Zeng, and B. Yu, "Hotspot detection via attention-based deep layout metric learning," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2020.
- [4] R. Liang, S. Nath, A. Rajaram, J. Hu, and H. Ren, "Buffermer: A generative ml framework for scalable buffering," in *IEEE/ACM Asia and South Pacific Design Automation Conference (ASPDAC)*, 2023, pp. 264–270.
- [5] S. Nath, G. Pradipta, C. Hu, T. Yang, B. Khailany, and H. Ren, "Transsizer: A novel transformer-based fast gate sizer," in *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*, 2022, pp. 1–9.
- [6] M. Liu, T.-D. Ene, R. Kirby, C. Cheng, N. Pinckney, R. Liang, J. Alben, H. Anand, S. Banerjee, I. Bayraktaroglu *et al.*, "Chipnemo: Domain-adapted llms for chip design," *arXiv preprint arXiv:2311.00176*, 2023.
- [7] Z. Pei, H.-L. Zhen, M. Yuan, Y. Huang, and B. Yu, "Betternv: Controlled verilog generation with discriminative guidance," *arXiv preprint arXiv:2402.03375*, 2024.
- [8] S. Liu, W. Fang, Y. Lu, Q. Zhang, H. Zhang, and Z. Xie, "Rtlcoder: Outperforming gpt-3.5 in design rtl generation with our open-source dataset and lightweight solution," *arXiv preprint arXiv:2312.08617*, 2023.
- [9] Y. Tsai, M. Liu, and H. Ren, "Rtlfixer: Automatically fixing rtl syntax errors with large language models," *arXiv preprint arXiv:2311.16543*, 2023.
- [10] M. Liu, N. Pinckney, B. Khailany, and H. Ren, "VerilogEval: Evaluating large language models for verilog code generation," in *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2023, pp. 1–8.
- [11] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.
- [12] "GPT-4o," 2024. [Online]. Available: <https://openai.com/index/hello-gpt-4o/>
- [13] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Conference on Neural Information Processing Systems (NIPS)*, vol. 36, 2024.
- [14] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [15] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, F. Weers *et al.*, "MM1: Methods, analysis & insights from multimodal llm pre-training," *arXiv preprint arXiv:2403.09611*, 2024.
- [16] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang, "Autogen: Enabling next-gen llm applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.
- [17] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv*, vol. abs/2306.14824, 2023. [Online]. Available: <https://arxiv.org/abs/2306.14824>
- [18] NVIDIA, "Neva: Nemo vision and language assistant," 2024, version 22B. [Online]. Available: <https://build.nvidia.com/nvidia/neva-22b>
- [19] M. Abidin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.
- [20] G. A. Team, "Paligemma: A lightweight open vision-language model (vlm)," *Google AI Blog*, May 2024. [Online]. Available: <https://blog.paperspace.com/paligemma-a-lightweight-open-vision-language-model-vlm/>
- [21] R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşlılar, "Fuyu-8b: A multimodal architecture for ai agents," *Adept AI*, 2024. [Online]. Available: <https://www.adapt.ai/fuyu-8b>
- [22] H. Ye, D.-A. Huang, Y. Lu, Z. Yu, W. Ping, A. Tao, J. Kautz, S. Han, D. Xu, P. Molchanov *et al.*, "X-vila: Cross-modality alignment for large language model," *arXiv preprint arXiv:2405.19335*, 2024.
- [23] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" *arXiv preprint arXiv:2307.06281*, 2023.
- [24] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *arXiv preprint arXiv:2308.02490*, 2023.
- [25] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang *et al.*, "Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," *arXiv preprint arXiv:2306.06687*, 2023.
- [26] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.
- [27] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao, "Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts," *arXiv preprint arXiv:2310.02255*, 2023.
- [28] J. Lin, H. Yin, W. Ping, P. Molchanov, M. Shoenybi, and S. Han, "Vila: On pre-training for visual language models," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 689–26 699.
- [29] "LLaMa-3.2," 2024. [Online]. Available: <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>
- [30] "Ollama," 2024. [Online]. Available: <https://ollama.com/>
- [31] "NVIDIA NIM," 2024. [Online]. Available: <https://www.nvidia.com/en-us/ai/>
- [32] "Azure OpenAI," 2024. [Online]. Available: <https://azure.microsoft.com/en-us/products/ai-services/openai-service>
- [33] "LLaMa-3," 2024. [Online]. Available: <https://llama.meta.com/llama3/>
- [34] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.
- [35] C. Liang *et al.*, "Fastchat," <https://github.com/lm-sys/FastChat>, 2023.
- [36] "Yi-34b," [Online]. Available: <https://github.com/01-ai/Yi>