Circuit-MLLM: Topological Logic-Guided Latent-Space Visual Reasoning for Circuit Schematic Understanding

Jinyuan Deng, Yuqi Jiang, Wenjing Huang, Xin Li, Qi Sun*, and Cheng Zhuo

Zhejiang University, Hangzhou, China
qisunchn@zju.edu.cn

Abstract. Through pre-training on extensive text and image datasets, current multi-modal large language models (MLLMs) achieve strong performance on general tasks. However, circuit schematics present a unique challenge for MLLMs due to their dense component layouts and distinct topological logic, demanding fine-grained structural parsing to extract the electrical semantics. To address this, we propose Circuit-MLLM, a multimodal reasoning framework that reformulates circuit topology analysis as a process of device localization, path tracing, and sequential reasoning within the latent space. We introduce a circuit knowledge mining mechanism that deeply aligns the model's latent representations with structurally rich features derived from multi-granularity circuit vision experts, enabling the model to effectively internalize topological semantics. Building upon these internalized semantics, we devise a topology-guided sequencing strategy that decouples reasoning from the rigid raster-scan order, enforcing stepwise inference along the circuit's topological logic in latent space. Across diverse circuit analysis tasks, Circuit-MLLM consistently outperforms strong baselines, notably achieving a 25% higher average score than GPT-5.1, which demonstrates the effectiveness of our framework in circuit schematic topology analysis. Code is publicly available at https://github.com/IC-Yuan/Circuit-MLLM.

Keywords: Multimodal Reasoning · Latent-space · Circuit Schematics

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) have shown strong performance on general vision–language understanding and reasoning [11, 12]. However, applying MLLMs to circuit schematics remains challenging due to their dense component layouts and intricate topological dependencies [20]. This limitation impedes the deployment of MLLMs in Electronic Design Automation (EDA). Consequently, empowering MLLMs with schematic parsing capabilities is pivotal for automating circuit analysis and accelerating chip design.

* Corresponding author.

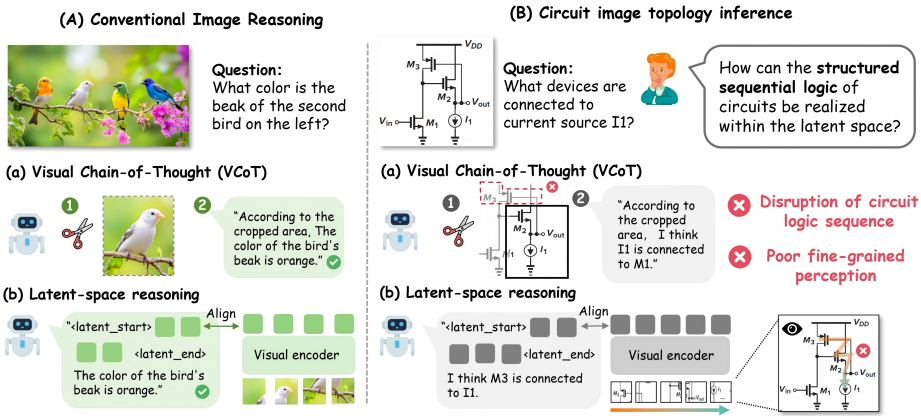

Fig. 1: Motivation of Circuit-MLLM. Existing Visual Chain-of-Thought methods often disrupt the topological integrity of circuits. Meanwhile, latent-space approaches, constrained by raster-order alignment and coarse-grained feature extraction, force the model to deviate from topological paths and result in a lack of fine-grained perception.

Traditional circuit analysis predominantly depends on rigid computer vision pipelines that employ rule-based algorithms for component and connection recognition [16]. Following this paradigm, several MLLM-based approaches for circuit understanding rely on specialized detectors as external front-ends for symbol localization [1], adopting a decoupled design that suffers from error propagation and inhibits the MLLM's inherent reasoning capabilities. In contrast, end-to-end approaches [22] bypass external detectors and typically rely on supervised fine-tuning (SFT) with schematic-to-structured language alignment. However, due to the lack of rich instance-level annotations, these models often rely on large-scale synthetic data, which limits their generalization to real-world distributions.

Recent advancements in MLLMs often employ Textual Chain-of-Thought to enhance reasoning, attempting to use textual descriptions to guide the model's attention toward specific image regions. For circuit schematics, however, describing topology and routing precisely in natural language is inherently brittle, frequently leading to semantic ambiguity and background distraction. Visual Chain-of-Thought (VCoT) [21] offers a novel perspective by explicitly invoking external tools such as image cropping. Nevertheless, these approaches [2, 13] typically adopt a region-centric "locate-judge" paradigm. This design is fundamentally ill-suited for circuit topology analysis, as discrete bounding boxes often fail to preserve the continuity of thin, long-range wires, resulting in fragmented connections and excessive background clutter.

In contrast, latent-space visual reasoning [4, 10, 19] offers an alternative by performing inference within hidden states, avoiding explicit bounding-box and text-based representations. This paradigm provides a distinct advantage for analyzing irregular topologies that are inherently difficult to describe. However, directly applying this latent paradigm to complex circuit topologies faces two

key challenges: **(1) Scarcity of fine-grained features:** Current latent-space approaches typically capture only coarse, macroscopic topological regions. These latent representations fail to encode the highly fine-grained connection details and dense component layouts essential for circuit topology analysis. **(2) Missing structured order:** Existing latent space alignment methods primarily enforce feature-level alignment between latent vectors and auxiliary image patches. Consequently, the sequence of latent representations defaults to a spatial, top-left to bottom-right raster-scan order. However, this scanning order fails to constrain the topology-driven viewpoint progression, resulting in a fundamental misalignment with the topology's intrinsic logic.

To overcome these challenges, we propose Circuit-MLLM, the first topological logic-guided latent-space visual reasoning framework dedicated to complex circuit topology analysis. First, to address the lack of fine-grained circuit feature perception, we introduce a **circuit knowledge mining mechanism** based on the latent space. By constructing circuit vision experts, we guide the model to deeply internalize and align complex circuit representations directly within the latent space during training. This mechanism prevents the model from merely focusing on superficial, coarse regional information, thereby endowing it with genuine, fine-grained circuit recognition capabilities. Furthermore, we propose a **topology-guided sequencing strategy.** This strategy mandates that the generated latent vectors not only accurately encode the topological regions but also strictly follow the intrinsic topological logic in their generation order by specifically progressing from the root component, along the wire topology, to the target component, thereby effectively simulating the topological viewpoint shifts of human engineers entirely within the latent space.

Our contributions are as follows:

1. We propose Circuit-MLLM, a novel latent-space visual reasoning paradigm for end-to-end multimodal analysis of circuit topologies. Unlike verbose Chain-of-Thought or external tool-chain methods, our framework directly bridges the semantic gap between spatial visual priors and rigorous circuit logic within the latent space, enabling intrinsic schematic comprehension.
2. We design a circuit knowledge mining mechanism based on latent space guided by a set of vision experts, which can generate highly discriminative fine-grained circuit latent representations without increasing inference overhead, thereby significantly improving topology recognition accuracy.
3. We propose a topology-guided sequencing strategy that decouples latent reasoning from rigid spatial raster-scanning. By dynamically aligning the inference trajectory with the inherent circuit structure, it endows the model with native, human-like topological tracing capabilities.
4. Extensive experiments reveal that Circuit-MLLM achieves a 25% higher average score than GPT-5.1 across diverse circuit analysis tasks, and outperforms latent-space visual baselines by up to 26% on topology reasoning tasks.

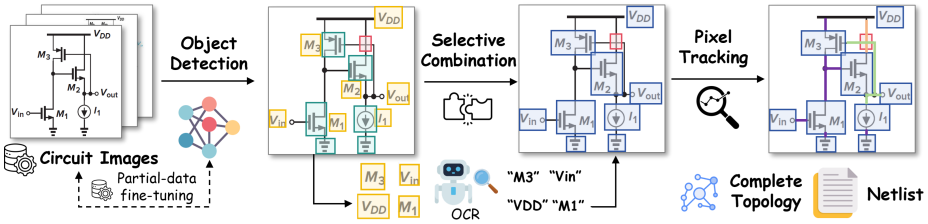

Fig. 2: Automated topology annotation workflow for circuit schematics

2 Preliminaries

2.1 Automated Circuit Schematic Topology Extraction

We have constructed an automated, fine-grained annotation workflow for circuit schematics as shown in Figure 2. First, we collected a total of 12,000 circuit schematics from multiple public datasets [1, 6, 15, 22], comprehensively covering analog, digital, and mixed-signal systems, making it the largest circuit schematic dataset currently available. Subsequently, we manually annotated a subset of the components, text, and junction nodes. A YOLO11 model [9] was then trained on this subset to enable automated detection and bounding box extraction across the entire dataset. Following this, we introduced an OCR framework [3] to perform text recognition and component matching, ensuring the accuracy of attribute assignments. Next, we employed pixel tracking techniques to reconstruct the topological dependencies of the circuits. Given the complexity of circuit topologies and connections, human expert intervention was incorporated to ensure rigorous accuracy. The resulting dataset provides netlist-style descriptions, visual grounding, and pixel-level regional masks. Simultaneously, we constructed latent-space reasoning QA pairs tailored to different question types, and introduced Circuit-MLLM-Bench to evaluate the model's capabilities in circuit topology analysis, as detailed in Section 4.

2.2 Related Works

Circuit Schematic Comprehension with MLLMs. Circuit schematics are a fundamental form of structured visual data in EDA, and their understanding has recently attracted growing attention in the MLLM community [17, 18]. Prior work mainly follows two paradigms. Tool-based visual reasoning methods (e.g., Masala CHAI [1]) rely on conventional vision components such as object detectors to produce structured annotations that guide MLLM reasoning. Text-conversion pipelines (e.g., MAPs [22]) first translate circuit schematics into structured netlists using supervised fine-tuned models or off-the-shelf parsers, and subsequently delegate reasoning to text-domain LLMs. However, both paradigms are bottlenecked by intermediate tools and conversions, suffering from error propagation and engineering overhead.

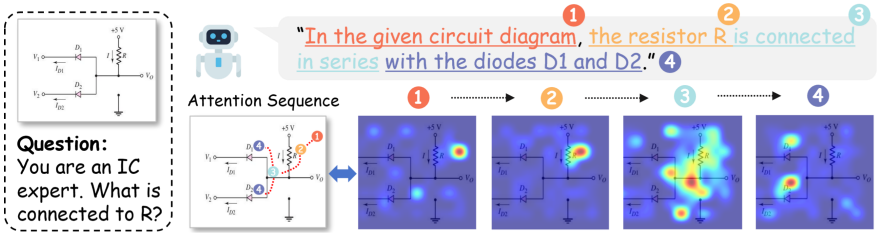

Fig. 3: Visualization of MLLM intrinsic attention flow in circuit topology analysis. The sequence illustrates native visual tracing from the root component, propagating along topological wire paths to terminal devices.

Multimodal Chain-of-Thought. Chain of Thought (CoT) prompting unlocked progressive reasoning in large language models via intermediate logical steps [5]. Recent studies extended CoT into multimodal settings. For instance, Vision R1 [8] deduces complex logic directly from images via multimodal reinforcement learning, while ICoT [21] interleaves attention-selected image crops with text tokens to improve visual question answering. Visual CoT [13] provides bounding box-grounded reasoning, enhancing spatial localization via explicit visual tokens. Others [7] employ external tools for visual evidence, augmenting multimodal CoT. However, these approaches encounter limitations in circuit schematics, where irregular topological regions are ill-suited for standard bounding boxes and challenging to trace via pure natural language.

Latent Visual Reasoning. Recent studies have introduced latent space reasoning methods [14] into multimodal domains. Specifically, these methods represent intermediate visual steps within the latent space rather than explicitly decoding them into images, thereby preventing the model from being constrained by rigid formats. LVR [10] and Mirage [19] pioneered the Latent Visual Reasoning paradigm, accomplishing latent reasoning by aligning auxiliary image features with vectors in the latent space. Building upon this, ILVR [4] introduced selective latent vector alignment alongside an interleaved multiple round latent space reasoning paradigm. Despite the unconstrained latent space being highly suitable for complex circuit topologies, current methods often overlook its intrinsic logical chains, causing reasoning misalignments in circuit topology analysis.

2.3 Visual Attribution of MLLMs in Circuit Analysis

To investigate how the model intrinsically engages in a native reasoning process to analyze circuit schematic topology without any explicit prompt guidance, we conducted a visual attribution of the MLLM's attention using the Qwen-2.5-VL 7B model. As depicted in Figure 3, we identified two key insights:

– **Insight 1: Topology-Guided Attention Flow.** When tackling circuit topology tasks, the MLLM exhibits a visual tracking pattern that mirrors circuit

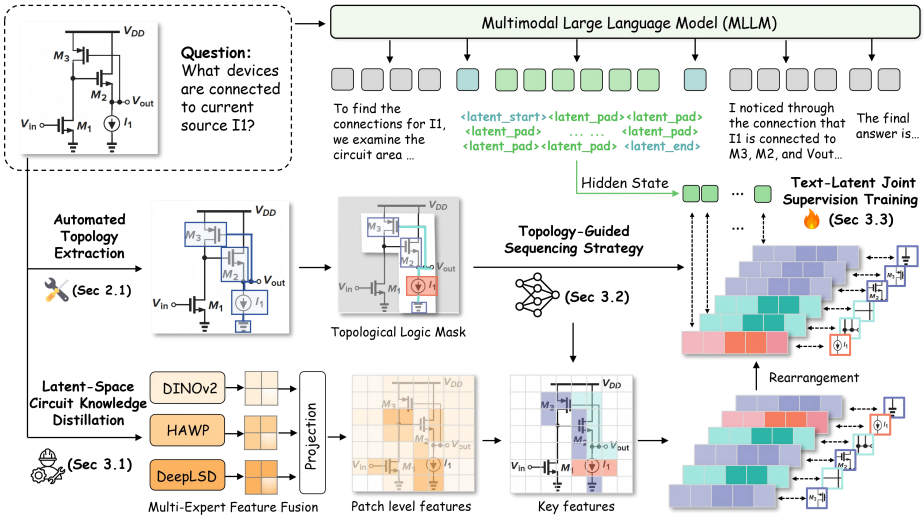

Fig. 4: Framework of Circuit-MLLM. (1) The circuit knowledge mining mechanism based on latent space provides the model with expert-level circuit representation vectors; (2) a Topology-Guided Sequencing Strategy directs latent-space reasoning to strictly follow the circuit's topological order; and (3) Text-Latent Joint Supervision Training unifies latent-space feature alignment and explicit text supervision.

connection logic. Rather than scanning components globally, the model initially anchors on a root component, subsequently traces strictly along topological wire paths, and ultimately shifts focus to downstream devices.

– **Insight 2: Concurrent Branch Tracking.** The MLLM demonstrates the capacity to simultaneously attend to multiple wire branches originating from a single node—a behavior likely facilitated by the multi-head attention mechanism. This parallel processing capability stands in contrast to human visual cognition, which typically favors serial, branch-by-branch analysis.

3 Methods

In this section, we introduce our Circuit-MLLM framework for sequentially ordered topological reasoning within the latent space. First, we present the latent-space circuit knowledge mining mechanism (Section 3.1) to internalize multi-expert visual capabilities without invoking external tools. Next, Section 3.2 details the topology-guided sequencing strategy, which strictly aligns the model's latent reasoning sequence with rigorous circuit logic. Finally, Section 3.3 introduces the text-latent joint supervision training to ensure alignment between internal implicit topological reasoning and explicit textual outputs.

3.1 Latent-Space Circuit Knowledge Mining Mechanism

While recent latent-space visual reasoning works show promising results, they primarily rely on generic visual encoders to extract basic image features. However, in specialized domains like circuit analysis, fine-grained visual comprehension (e.g., component topologies) is paramount. To acquire these details, conventional methods often integrate supplementary network modules or explicitly invoke external vision tools during inference. This disjointed approach easily introduces cascading errors and increases inference latency.

To overcome these limitations, we propose a Latent-Space Circuit Knowledge Mining Mechanism. Our core philosophy is to abandon inference-time reliance on external tools, enabling the Circuit-MLLM to intrinsically learn and assimilate multi-expert visual features directly within its latent space during training. To this end, we design a Multi-Expert Feature Fusion module that serves as the source of knowledge guidance. For a given circuit image I, we deploy three heterogeneous vision experts to extract complementary structural and semantic priors: HAWP for holistic wireframe and junction detection ($\mathcal{F}_{\text{HAWP}}$), DeepLSD for fine-grained topological line segment extraction ($\mathcal{F}_{\text{LSD}}$), and DINOv2 for robust patch-level semantic representations ($\mathcal{F}_{\text{DINO}}$).

Due to architectural disparities among these experts, the extracted feature maps vary in spatial resolutions and channel dimensions. To construct a unified representation, we align them to a target spatial grid (H_s, W_s) determined by the Circuit-MLLM's visual encoder. Let $\Phi_{\text{align}}(\cdot)$ denote the bilinear interpolation operator mapping a tensor to this target resolution. The aligned features are concatenated along the channel dimension to form a dense expert representation $\mathcal{F}_{\text{ext}}$:

$$\mathcal{F}_{\text{ext}} = [\Phi_{\text{align}}(\mathcal{F}_{\text{HAWP}}), \Phi_{\text{align}}(\mathcal{F}_{\text{LSD}}), \Phi_{\text{align}}(\mathcal{F}_{\text{DINO}})] \in \mathbb{R}^{(C_1+C_2+C_3)\times H_s \times W_s}. \tag{1}$$

Next, we flatten the spatial dimensions and employ a learnable projection matrix $\mathbf{W}_{\text{proj}}$ to project the concatenated heterogeneous features into the identical latent dimension d of the Circuit-MLLM. Let $\mathbf{E}_{\text{orig}} \in \mathbb{R}^{N\times d}$ denote the initial visual tokens extracted by the Circuit-MLLM's original visual encoder, where $N = H_s \times W_s$. To imbue these initial tokens with circuit-specific priors, we concatenate them and pass them through a fusion layer $\mathbf{W}_{\text{fuse}}$ to obtain the knowledge-enriched feature pool $\mathbf{E}_{\text{fused}}$:

$$\mathbf{E}_{\text{expert}} = \text{Flatten}(\mathcal{F}_{\text{ext}})\mathbf{W}_{\text{proj}}, \tag{2}$$

$$\mathbf{E}_{\text{fused}} = [\mathbf{E}_{\text{orig}}, \mathbf{E}_{\text{expert}}]\,\mathbf{W}_{\text{fuse}}. \tag{3}$$

The fused feature pool $\mathbf{E}_{\text{fused}}$ provides a set of candidate alignment targets for Section 3.2 and Section 3.3, forcing the MLLM to internalize expert-enhanced visual semantics within its latent space.

3.2 Topology-Guided Sequencing Strategy

Motivated by the unique attention behaviors observed in our visual attribution analysis in Section 2.3, we introduce the **Topology-Guided Sequencing**

Strategy. This latent-space reasoning paradigm explicitly guides the model to follow the sequential perspective of topological logic, aligning its inherent tracking capabilities with the rigorous demands of circuit analysis.

Instead of relying on coarse bounding boxes that introduce background noise, we utilize the automated topology extraction workflow (Section 2.1) to obtain exact pixel-level localization for each topological query. Guided by the visual attention shifts of Circuit-MLLM and human domain expertise, we formulate a pixel-wise sequencing map, termed the **Topological Logic Mask** $S(p)$:

$$S(p) = \begin{cases} 0, & \text{if } p \in \Omega_{\text{bg}} \\ 1, & \text{if } p \in \Omega_{\text{root}} \\ 2 + \frac{\mathcal{D}_{\text{BFS}}(p, \Omega_{\text{root}})}{\max_{q \in \Omega_{\text{wire}}} \mathcal{D}_{\text{BFS}}(q, \Omega_{\text{root}})}, & \text{if } p \in \Omega_{\text{wire}} \\ 3 + \gamma_i, & \text{if } p \in \Omega_{\text{conn}}^{(i)} \end{cases} \tag{4}$$

where $\mathcal{D}_{\text{BFS}}$ denotes the topological distance calculated via Breadth-First Search, capturing the "near-to-far" principle of visual attention. Meanwhile,$\gamma_i \sim \mathcal{U}(0, 1)$ assigns a distinct continuous value to differentiate connected devices.

To align this pixel-level mask with the patch-level features $\mathbf{E}_{\text{fused}}$, we apply Adaptive Max Pooling, yielding the patch-level Topological Logic Mask $\hat{S}_{i,j} = \max_{p \in \mathcal{P}_{i,j}} S(p)$. This acts as a conservative filter to preserve highly localized structures like thin wires. Crucially, after filtering out background features ($\hat{S}_{i,j} > 0$), we explicitly inject topological logic by defining a permutation π on the valid feature indices $\mathcal{V}$. This permutation sorts the indices such that their corresponding mask values are monotonically increasing: $\hat{S}_{\pi(1)} \leq \hat{S}_{\pi(2)} \leq \cdots \leq \hat{S}_{\pi(|\mathcal{V}|)}$. This seamlessly transforms the unordered 2D spatial features into a strictly topologically-ordered 1D sequence:

$$\tilde{\mathbf{E}}_{\text{fused}} = [\mathbf{E}_{\text{fused},\pi(1)}, \mathbf{E}_{\text{fused},\pi(2)}, \ldots, \mathbf{E}_{\text{fused},\pi(|\mathcal{V}|)}]. \tag{5}$$

To accommodate the MLLM's fixed latent capacity k, a 1D Adaptive Average Pooling compresses this sequence into a fixed-length target $\mathbf{E}^* \in \mathbb{R}^{k \times d}$:

$$\mathbf{E}^*_m = \frac{1}{|B_m|} \sum_{t \in B_m} \tilde{\mathbf{E}}_{\text{fused},t}, \quad m \in \{1, 2, \ldots, k\}. \tag{6}$$

This ordered, expert-enriched sequence serves as the target latent vectors for knowledge guidance. During the autoregressive generation process, we employ a **Sequential Topological-Knowledge Alignment Loss** ($\mathcal{L}_{\text{topo}}$) to optimize the model by minimizing the cosine distance between the hidden state h_m at the final layer and the target latent vectors:

$$\mathcal{L}_{\text{topo}} = \frac{1}{k} \sum_{m=1}^{k} \left(1 - \cos(h_m, \mathbf{E}^*_m)\right). \tag{7}$$

This mechanism effectively constrains the model to traverse predefined topological trajectories, thereby anchoring its latent representations within a logically structured and dynamically sequenced reasoning space.

3.3 Text-Latent Joint Supervision Training

While the latent tokens are supervised by the unified topological knowledge alignment objective $\mathcal{L}_{\text{topo}}$, the surrounding text tokens are optimized via a standard autoregressive cross-entropy loss. Let $\boldsymbol{x}$ denote the input query and circuit image, and $\boldsymbol{o}_{\text{pre}}$ / $\boldsymbol{o}_{\text{post}}$ denote the text segments respectively preceding and succeeding the generated latent tokens $\{h_m\}_{m=1}^{k}$. The textual loss $\mathcal{L}_{\text{text}}$ is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{text}} = &\sum_{i=1}^{|\boldsymbol{o}_{\text{pre}}|} \ell_{\text{CE}}\left(\boldsymbol{o}_{\text{pre},i}, P_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre},<i})\right) \\ &+ \sum_{i=1}^{|\boldsymbol{o}_{\text{post}}|} \ell_{\text{CE}}\left(\boldsymbol{o}_{\text{post},i}, P_\theta(\boldsymbol{x}, \boldsymbol{o}_{\text{pre}}, \{h_m\}_{m=1}^{k}, \boldsymbol{o}_{\text{post},<i})\right),\end{aligned} \tag{8}$$

where $P_\theta(\cdot)$ denotes the next-token prediction probability of the Circuit-MLLM. The overall training objective seamlessly combines both terms:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{topo}} + \mathcal{L}_{\text{text}}, \tag{9}$$

where λ is a balancing coefficient. This joint optimization anchors the latent tokens in the topologically-ordered expert visual space while weaving them naturally into the model's explicit textual reasoning.

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate our method on Circuit-MLLM-bench, a custom large-scale circuit benchmark. This benchmark is constructed via the Automated Circuit Schematic Topology Extraction pipeline detailed in Section 2.1 and has been rigorously verified by human experts. The benchmark evaluates models across two primary categories: topological analysis tasks, consisting of Connection Judgment and Connection Identification; and conventional detection tasks, which include Total Count, Type Count, and Element Class. Detailed construction procedures are provided in **??**.

Data Synthesis. Driven by our automated topology extraction workflow, we generate question-answer (QA) pairs for each circuit image across the five tasks, alongside auxiliary images and pixel-level masks that delineate semantic regions and govern latent topological sequencing. We sample 5,000 instruction-tuning instances distributed by task complexity: Connection Identification (30%), Connection Judgment (25%), Element Class (15%), Total Count (15%), and Type Count (15%). To prevent data leakage, training images are strictly mutually exclusive from the Circuit-MLLM-bench evaluation set.

Baselines. We compare our approach against four baseline categories: (1) General-Purpose MLLMs: Mainstream models, encompassing closed-source models (e.g.,

Table 1: Performance comparison across all task categories. "Acc.", "F1", and "EMR" denote Accuracy, F1-score, and Exact Match Ratio, respectively. "Acc." is evaluated on single-choice and counting tasks, whereas "F1" and"EMR" are applied to multiple-response tasks. "Avg." represents the overall average score.

Models	Size	Circuit-MLLM-bench						
		Total Count	Type Count	Element Class	Conn.Judge	Conn.Identi		Avg
		Acc.↑	Acc.↑	Acc.↑	Acc.↑	F1↑	EMR.↑	
GPT 5.1	/	43.60	60.62	96.49	61.11	76.08	30.20	61.35
GPT-4o	/	42.01	60.08	97.40	58.50	78.68	36.30	62.16
Claude-4-Sonnet	/	44.91	58.23	95.13	57.13	69.54	24.80	58.17
Doubao-1.5-vision-pro	/	40.21	65.00	91.35	60.36	74.08	24.40	59.23
deepseek-vl2	27B	24.13	43.80	77.62	56.56	59.61	17.00	46.45
GLM-4.5V	106B	58.62	59.50	90.08	58.59	79.37	40.00	64.36
Qwen2.5-VL	7B	21.06	48.40	83.55	55.92	65.53	15.20	48.28
	32B	26.77	57.60	88.02	60.14	68.77	14.80	52.68
	72B	29.95	58.10	90.60	57.31	72.73	24.60	55.55
Qwen3-VL	8B	39.47	56.00	90.30	57.52	76.86	38.30	59.74
	32B	49.63	60.00	92.95	58.59	82.02	42.90	64.35
Masala-CHAI	7B	27.72	48.10	93.10	51.00	66.35	19.10	50.88
Circuit-MLLM	7B	**75.87**	**65.10**	**97.40**	**72.60**	**88.90**	**57.30**	**76.20**

GPT 5.1,GPT-4o, Claude-4-Sonnet, Doubao-1.5-vision-pro) and open-source models (e.g., Qwen 2.5 and Qwen 3 series, DeepSeek-VL2, and GLM-4.5V). (2) Latent-Space Visual Reasoning: Mirage [19] and ILVR [4].(3) Tool-Augmented Circuit MLLMs: Masala CHAI [1], representing methods that explicitly invoke external tools to modify images prior to analysis. (4) Direct Fine-Tuning: Supervised Fine-Tuning (SFT) and SFT combined with Group Relative Policy Optimization (SFT+GRPO), representing models trained directly on our dataset without utilizing the latent-space mechanism.

Implementation Details. All experiments employ Qwen2.5-VL 7B as the foundational base model. Our framework is implemented using PyTorch and trained on 8 NVIDIA A100 GPUs. We conduct supervised fine-tuning for 15 epochs, utilizing a batch size of 8 alongside a cosine learning rate scheduler with an initial learning rate of 1e-5 across both stages. The latent token size is set to $k = 4$, and the loss balancing coefficient is configured to $\gamma = 0.6$.

4.2 Experimental Results

Results on all task categories. Table 1 presents the comparative evaluation results of various open-source and closed-source models across all task categories on Circuit-MLLM-Bench. In terms of average performance, Circuit-MLLM outperforms all existing baseline models, achieving a significant 57.8% relative improvement over its backbone, Qwen-2.5-VL-7B, and an approximate 22.5% improvement over the leading model GPT-4o. This underscores the effectiveness of Circuit-MLLM in solving complex circuit-related problems. Furthermore, although Masala-CHAI, a method utilizing external visual tools, holds a slight

Table 2: Performance comparison on topological analysis tasks. The evaluation encompasses two tasks: Connection Judgment and Connection Identification. The difficulty levels (Easy, Medium, and Hard) are determined by the number of connected components and the total number of components within the circuit diagram.

Models	Methods		Connection Judge.			Connection Identification						Avg
			Easy	Medium	Hard	Easy		Medium		Hard		
			Acc.↑	Acc.↑	Acc.↑	F1↑	EMR.↑	F1↑	EMR.↑	F1↑	EMR.↑	
	SFT	GRPO										
Qwen2.5-VL 7B			59.29	51.56	41.52	75.24	44.01	61.70	7.20	61.54	3.78	45.09
	✓		76.70	63.28	56.68	87.86	69.46	89.92	59.20	80.40	26.12	67.74
	✓	✓	76.70	62.76	57.40	87.86	69.76	89.85	58.13	80.53	26.80	67.75
	Masala-CHAI		55.16	53.91	41.88	75.70	45.21	61.87	6.93	61.46	4.47	45.18
Latent Reasoning 7B	Mirage		77.88	66.15	64.98	90.41	72.16	89.48	58.67	81.88	26.80	69.82
	ILVR		79.06	63.28	63.90	89.74	72.46	88.98	55.73	79.47	22.68	68.37
	Circuit-MLLM		**81.42**	**70.31**	**64.98**	**91.44**	**76.35**	**91.10**	**62.67**	**83.54**	**28.52**	**72.26**

edge over Qwen-2.5-VL-7B in total count, element class, and connection identification, its overall gain is merely 4%, indicating the limitations of relying purely on external tools. Most importantly, on the highly challenging connection identification task, our method achieves a breakthrough in the Exact Match Ratio (EMR), reaching approximately three times that of the backbone model, thereby demonstrating its robust capability in parsing complex topologies.

Results on topological analysis tasks. To evaluate Circuit-MLLM in topological analysis, we fine-tuned the latent-space visual baselines, Mirage [19] and ILVR [4], using our latent-space dataset with the same latent size of 4 to ensure a fair comparison. As shown in Table 2, Circuit-MLLM consistently outperforms both Mirage and ILVR across all difficulty levels in Connection Judgment and Connection Identification. The average score increases by approximately 5%, demonstrating the overall effectiveness of our method. Notably, on the Hard difficulty of Connection Identification, our EMR exhibits a substantial 25% improvement over ILVR. Furthermore, Masala-CHAI performs almost identically to the backbone model, Qwen-2.5-VL-7B, on topological tasks, suggesting that explicitly modifying the original image yields marginal benefits. Compared to the SFT+GRPO baseline, Circuit-MLLM achieves an overall improvement of about 7%, with maximum gains reaching 13% on specific tasks. Finally, latent-space visual methods generally surpass the direct SFT baseline (without latent mechanisms) across topological tasks. This consistent superiority strongly proves the validity of conducting topological analysis directly within the latent space.

4.3 Ablation Study

Effectiveness of the Proposed Method. As shown in Table 3, the baseline latent reasoning model achieves an average score of 69.85. Introducing the Topology-Guided Sequencing Strategy alone improves the average score to 70.61,

Table 3: Ablation Study of our methods. "Seq." and "Expert" refer to our two core components: the Topology-Guided Sequencing Strategy and Latent-Space Circuit Knowledge Distillation, respectively.

Models	Method		Connection Judge.			Connection Identification						Avg
	Seq.	Expert	Easy	Medium	Hard	Easy		Medium		Hard		
			Acc.↑	Acc.↑	Acc.↑	F1↑	EMR.↑	F1↑	EMR.↑	F1↑	EMR.↑	
Latent Reasoning 7B			77.88	65.10	62.45	89.97	73.05	90.60	61.87	82.28	25.46	69.85
	✓		79.06	66.15	64.26	89.99	73.05	90.98	62.40	**83.86**	25.77	70.61
		✓	77.88	65.62	62.50	90.64	73.35	**91.22**	**63.47**	83.83	27.15	70.64
	✓	✓	**81.42**	**70.31**	**64.98**	**91.44**	**76.35**	91.10	62.67	83.54	**28.52**	**72.26**

Table 4: Ablation Study of Latent Size. The latent size denotes the total number of latent vectors generated during the inference phase. Across all experiments in this section, the loss balancing coefficient is uniformly configured to $\gamma = 0.6$.

Models	Latent Size	Total Count	Type Count	Element Class	Conn.Judge	Conn. Ident		Avg
		Acc.↑	Acc.↑	Acc.↑	Acc.↑	F1↑	EMR.↑	
Latent Reasoning 7B	2	74.05	66.60	96.20	67.70	87.52	54.50	74.42
	4	**75.87**	65.10	**97.40**	**72.60**	88.92	**57.30**	**76.20**
	6	73.75	66.70	95.10	67.80	87.49	54.30	74.19
	8	75.13	**67.80**	96.50	67.70	**89.12**	56.30	75.43

bringing consistent gains in Connection Judgment across all difficulty levels. Conversely, incorporating only the Latent-Space Circuit Knowledge mining mechanism raises the average score to 70.64, showing particular effectiveness in the Connection Identification task. When both components are integrated, Circuit-MLLM achieves the highest overall average score of 72.26, a 2.41 improvement over the baseline. The full model demonstrates superior performance across all Connection Judgment tasks and attains the highest EMR (28.52) on the Hard difficulty of Connection Identification. These quantitative results confirm that the two components are complementary, effectively enhancing the model's topological analysis capabilities within the latent space.

Selection of latent size. Table 4 presents the performance of Circuit-MLLM on the Circuit-MLLM-Bench across different latent sizes. In terms of the overall score, a latent size of 4 yields the best results, representing the most balanced choice. Furthermore, this indicates that 4 latent pads are sufficient to encapsulate the reasoning process for conventional circuit problems.

Effectiveness of different circuit visual experts. Table 5 presents the performance of our three circuit visual experts, DINOv2, HAWP, and DeepLSD, across five circuit tasks. Employing all experts simultaneously yields the highest average score, demonstrating their collective effectiveness in circuit problem analysis. Furthermore, the individual contributions of each expert vary significantly across different problem types. For instance, utilizing only DINOv2 leads to strong performance on counting tasks, such as Total Count and Type Count,

Table 5: Ablation study of circuit experts. We evaluate the impact of different expert models, including DINOv2 , HAWP (Holistically-Attracted Wireframe Parsing, and DeepLSD (Deep Line Segment Detection). In all configurations, the latent size is set to 6, and the loss balancing coefficient is uniformly configured to $\gamma = 0.6$.

Circuit Experts			Total Count	Type Count	Element Class	Conn.Judge	Conn. Ident		Avg
DINO	HAWP	DeepLSD	Acc.↑	Acc.↑	Acc.↑	Acc.↑	F1↑	EMR.↑	
✓			75.23	**66.80**	94.50	69.00	88.60	56.80	75.16
	✓		71.11	65.90	95.70	70.60	**89.16**	57.20	74.95
✓	✓		74.49	66.50	96.40	71.00	88.16	54.80	75.23
✓	✓	✓	**75.87**	65.10	**97.40**	**72.60**	88.92	**57.30**	**76.20**

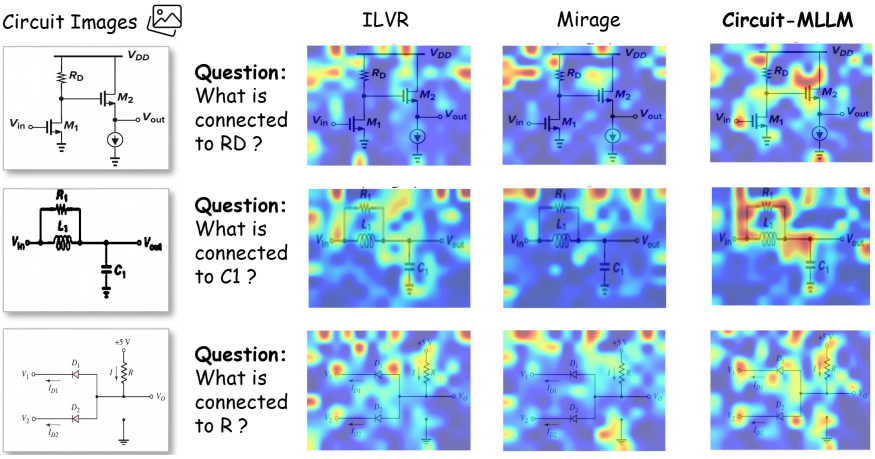

Fig. 5: Visual comparisons in the latent space. The latent features are captured prior to the generation of each <latent pad> token. Compared to Mirage and ILVR, our Circuit-MLLM achieves both higher fine-grained resolution and precise localization.

highlighting its precision in general object recognition; however, it struggles with connection-related problems. Conversely, relying solely on HAWP yields better results on Conn. Judge and Conn. Ident., verifying that visual experts specialized in wire detection are indeed highly effective for topology-related tasks.

Additional Results. Further experimental results, including coefficient γ selection and Amsbench [17] evaluations, are provided in **??**.

5 Analysis

Latent Space Attention Analysis. To investigate the visual regions focused on by the MLLM in the latent space, we extract the latent features from the last layer prior to the generation of each <latent pad> token. These features are utilized as latent vectors and mapped to the input image to generate heatmaps as shown in Figure 5. Observations indicate that when the input text queries

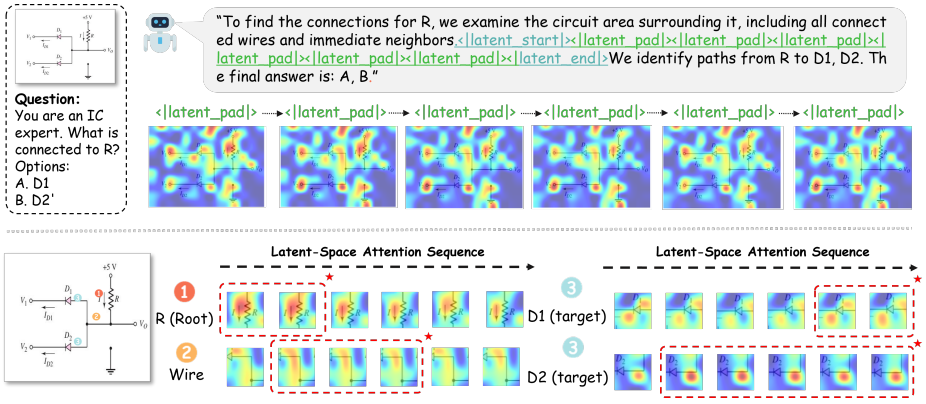

Fig. 6: Dynamic visual shifts within the latent space. We demonstrate this using latent size of 6. **Top:** Global heatmaps of the penultimate layer's latent representations overlaid on the input image. **Bottom:** Localized heatmaps focusing on critical regions for topological reasoning, including the root device, wire, and target device.

specific components, our Circuit-MLLM accurately focuses on the root component, adjacent wires, and the final target component within the latent space. In contrast, Mirage and ILVR struggle to localize the regions of interest and exhibit lower fine-grained resolution, making it difficult to capture minute details such as wires. This demonstrates that our proposed Topological Logic Mask and Visual Expert effectively assist the model in capturing these crucial circuit topology features. Furthermore, we note that the latent space naturally avoids noise components and blank areas, proving that the latent space visual approach yields superior performance compared to direct image cropping.

Latent Space Visual Shift Analysis. To verify whether Circuit-MLLM achieves dynamic visual shifts in the latent space, we extracted six consecutive latent feature vectors as shown in the Top of Figure 6. We specifically focus on the critical regions for topological reasoning: the root device, the wire, and the target device. As illustrated in the Bottom of Figure 6, a temporal progression of visual attention is observed: the model initially focuses heavily on the root device, followed by a gradual decay of attention in this area; the visual focus then shifts along the connecting wires, ultimately concentrating on the target device. This trajectory aligns with human analytical workflow, proving Circuit-MLLM performs accurate topological reasoning entirely within the latent space.

6 Conclusion

In this paper, we proposed Circuit-MLLM, the first topological logic-guided latent-space visual reasoning framework dedicated to complex circuit topology analysis. To address the inherent irregularity of topological regions in circuit diagrams, we pioneered the embedding of the topological reasoning process into the

latent space. Extensive experiments demonstrate that by incorporating the circuit knowledge mining mechanism and the topology-guided sequencing strategy, Circuit-MLLM accurately focuses on the intended topological regions, leading to significant performance improvements in complex topological analysis tasks.

Acknowledgements

This work was supported by NSFC (Grant No. U25A20485, W2412034).

References

1. Bhandari, J., Bhat, V., He, Y., Rahmani, H., Garg, S., Karri, R.: Masala-chai: A large-scale spice netlist dataset for analog circuits by harnessing ai. arXiv preprint arXiv:2411.14299 (2024)
2. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1254–1262 (2024)
3. Cui, C., Sun, T., Liang, S., Gao, T., Zhang, Z., Liu, J., Wang, X., Zhou, C., Liu, H., Lin, M., et al.: Paddleocr-vl: Boosting multilingual document parsing via a 0.9 b ultra-compact vision-language model. arXiv preprint arXiv:2510.14528 (2025)
4. Dong, S., Wang, S., Liu, X., Li, C., Hou, H., Wei, Z.: Interleaved latent visual reasoning with selective perceptual modeling. arXiv preprint arXiv:2512.05665 (2025)
5. Feng, G., Zhang, B., Gu, Y., Ye, H., He, D., Wang, L.: Towards revealing the mystery behind chain of thought: a theoretical perspective. Advances in Neural Information Processing Systems **36**, 70757–70798 (2023)
6. Gao, J., Cao, W., Yang, J., Zhang, X.: Analoggenie: A generative engine for automatic discovery of analog circuit topologies. arXiv preprint arXiv:2503.00205 (2025)
7. Hu, Y., Shi, W., Fu, X., Roth, D., Ostendorf, M., Zettlemoyer, L., Smith, N.A., Krishna, R.: Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. Advances in Neural Information Processing Systems **37**, 139348–139379 (2024)
8. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
9. Jocher, G., Qiu, J.: Ultralytics yolo11 (2024), `https://github.com/ultralytics/ultralytics`
10. Li, B., Sun, X., Liu, J., Wang, Z., Wu, J., Yu, X., Chen, H., Barsoum, E., Chen, M., Liu, Z.: Latent visual reasoning. arXiv preprint arXiv:2509.24251 (2025)
11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
12. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
13. Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. Advances in Neural Information Processing Systems **37**, 8612–8642 (2024)

14. Shen, Z., Yan, H., Zhang, L., Hu, Z., Du, Y., He, Y.: Codi: Compressing chain-of-thought into continuous space via self-distillation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 677–693 (2025)
15. Shi, Y., Tao, Z., Gao, Y., Huang, L., Wang, H., Yu, Z., Lin, T.J., He, L.: Amsnet 2.0: A large ams database with ai segmentation for net detection. arXiv preprint arXiv:2505.09155 (2025)
16. Shi, Y., Tao, Z., Gao, Y., Zhou, T., Chang, C., Wang, Y., Chen, B., Zhang, G., Liu, A., Yu, Z., et al.: Amsnet-kg: A netlist dataset for llm-based ams circuit auto-design using knowledge graph rag. ACM Transactions on Design Automation of Electronic Systems (2024)
17. Shi, Y., Zhang, Z., Wang, H., Tao, Z., Li, Z., Chen, B., Wang, Y., Yu, Z., Lin, T.J., He, L.: Amsbench: A comprehensive benchmark for evaluating mllm capabilities in ams circuits. arXiv preprint arXiv:2505.24138 (2025)
18. Xiang, K., Li, H., Zhang, T.J., Huang, Y., Liu, Z., Qu, P., He, J., Chen, J., Yuan, Y.J., Han, J., et al.: Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. arXiv preprint arXiv:2505.19099 (2025)
19. Yang, Z., Yu, X., Chen, D., Shen, M., Gan, C.: Machine mental imagery: Empower multimodal reasoning with latent visual tokens. arXiv preprint arXiv:2506.17218 (2025)
20. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 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. pp. 9556–9567 (2024)
21. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
22. Zhu, E., Liu, Y., Zhang, Z., Li, X., Zhou, J., Yu, X., Huang, M., Wang, H.: Maps: Advancing multi-modal reasoning in expert-level physical science. arXiv preprint arXiv:2501.10768 (2025)