VG-CoT: Towards Trustworthy Visual Reasoning via Grounded Chain-of-Thought

Byeonggeuk Lim[1], Kyeonghyun Kim[2], Jungmin Yun[2] and YoungBin Kim[1, 2]
Graduate School of Advanced Imaging Science, Multimedia & Film, Chung-Ang University[1]
Department of Artificial Intelligence, Chung-Ang University[2]
{banggeuk, khyun8072, cocoro357, ybkim85}@cau.ac.kr

Abstract

The advancement of Large Vision-Language Models (LVLMs) requires precise local region-based reasoning that faithfully grounds the model's logic in actual visual evidence. However, existing datasets face limitations in scalability due to extensive manual annotation and lack of explicit alignment between multi-step reasoning and corresponding image regions, which constrains the evaluation of model trustworthiness. To address these challenges, we propose the Visual Grounding Chain-of-Thought (VG-CoT) dataset, which explicitly links each reasoning step to real visual evidence within the image through a fully automated three-stage pipeline. The pipeline first extracts object- and text-level visual evidence using state-of-the-art detection and OCR models, then generates step-by-step grounded reasoning with GPT-4o, and finally refines the grounding through a rationale-driven open-set detection process. In addition, we introduce a new benchmark that comprehensively evaluates LVLMs reasoning across three complementary dimensions: Rationale Quality, Answer Accuracy, and Reasoning-Answer Alignment. Experiments with representative LVLMs, including LLaVA-1.5 and Qwen2-VL, demonstrate consistent improvements on most evaluation metrics, confirming that VG-CoT effectively enhances trustworthy, evidence-based reasoning while maintaining scalable and cost-efficient dataset construction. The dataset and code will be released publicly upon acceptance to facilitate further research.

Keywords: LVLMs Reasoning, Visual Grounded Rationale Dataset, Reasoning Evaluation

1. Introduction

The development of Large Language Models (LLMs) has recently led to the rise of Large Vision-Language Models (LVLMs), which simultaneously understand visual and linguistic information (Zhang et al., 2024; Yin et al., 2024). LVLMs have demonstrated outstanding performance in comprehensive, image-level understanding, and the focus of research is gradually expanding toward the ability for precise understanding of local regions within an image (Dai et al., 2023; Liu et al., 2024; Zhu et al.; Wang et al., 2024). Specifically, the ability to accurately identify specific regions within an image and interpret their spatial and semantic relationships is emerging as a core requirement for performing advanced vision-language tasks such as autonomous driving, robotics, and medical image analysis (Li et al., 2024; Zitkovich et al., 2023; Li et al., 2023).

However, this local region-based reasoning ability is highly dependent on the quality and composition of the training dataset (Wu et al., 2023; Chang et al., 2024). Currently, most widely used vision-language datasets primarily focus on evaluating comprehensive, image-level understanding, which limits models from learning fine-grained reasoning based on the detailed attributes of individual objects or the relationships between objects (Chen et al., 2024b; Zhou et al., 2025; Antol et al., 2015; Hudson and Manning, 2019). Although some studies have proposed datasets that include local region information, they still rely on large-scale manual annotation, making it difficult to avoid the constraints of cost and scalability (Yu et al., 2016; Zellers et al., 2019; Krishna et al., 2017). This manual construction method presents a more pronounced limitation, especially in tasks that require complex visual relationships or multi-step reasoning. This suggests the need for a dataset that can support the learning of complex, local region-based reasoning in a scalable and efficient manner.

Existing multimodal Chain-of-Thought (CoT) datasets designed to support multi-step reasoning suffer from a structural limitation: the reasoning process is not explicitly linked to specific visual evidence within the actual image (Lu et al., 2022; Yu et al., 2016). In other words, because the location information (spatial coordinates) of the objects or text within the image that serve as the rationale for the inference is not provided, it is difficult to verify which areas of the image the model's reasoning was based on. This absence of visual evidence leads to a structural limitation in model evaluation regarding whether the model derived the correct answer using the right evidence as its rationale, causing existing benchmarks to only assess the accuracy of the final answer (Wu et al., 2023; Chang et al., 2024). Therefore, a dataset where the entire process of reasoning is explicitly aligned with real visual evidence is essential for verifying whether

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11108–11118
11–16 May 2026.

the logic presented by the model is faithful to the actual image evidence (Chang et al., 2024; Qiu et al., 2024). This also provides the foundation for a new benchmark capable of systematically measuring not only the accuracy of the final answer but also the accuracy of the rationale and the logical soundness of the reasoning process.

Based on this necessity, this study proposes the Visual Grounding Chain-of-Thought (VG-CoT) dataset. VG-CoT is a dataset constructed by explicitly aligning visual evidence from an image with the reasoning process across various tasks, achieving both efficiency and scalability through an automated three-stage pipeline. (i) First, initial visual evidence is extracted using state-of-the-art object detection and OCR models. (ii) Based on this, step-by-step reasoning is generated using GPT-4o. (iii) Finally, using the generated reasoning process as a clue, visual evidence is more precisely captured through a text-based object detection model. This automated pipeline effectively solves the cost issue associated with manual annotation while contributing to the improvement of the quality of the reasoning process and the reliability of rationale alignment.

Furthermore, this study proposes a new benchmark that goes beyond the existing evaluation methods centered on answer accuracy. This benchmark comprehensively measures Rationale Quality, Answer Accuracy, and Reasoning-Answer Alignment, which represents the relationship between the two metrics. Specifically, it utilizes GPT-4o (Achiam et al., 2023) as an evaluator to precisely assess the reasoning process using detailed metrics: logical coherence, reasoning completeness, and visual evidence utilization. This allows for an in-depth verification of whether the model has reasoned logically based on credible visual evidence, going beyond simply checking if the model provided the correct answer.

The key contributions of this study are summarized as follows:

1. We propose the VG-CoT dataset, which explicitly aligns the reasoning process with real visual evidence within the image through an automated three-stage pipeline.

2. We establish a new benchmark that comprehensively measures Rationale Quality and Reasoning–Answer Alignment, going beyond simple answer accuracy.

3. We evaluate representative LVLMs using VG-CoT to analyze their capability in utilizing visual evidence, providing insights into the future direction of LVLMs research.

2. Related Work

2.1. Datasets for LVLMs Reasoning

Early datasets for LVLMs primarily focused on image-wide understanding, which limited their ability to train models for fine-grained, local region-based reasoning (Lin et al., 2014; Antol et al., 2015; Hudson and Manning, 2019). To overcome this limitation, datasets that explicitly include local region information were proposed, but most relied on manual annotation, a method that requires immense cost and time, thereby restricting data diversity and scalability (Krishna et al., 2017; Peng et al., 2024; Yu et al., 2016). Consequently, there emerged a need for datasets that overcome the limitations of manual annotation and require scalable yet detailed visual understanding.

In line with this necessity, studies have been conducted to encourage models to perform complex reasoning processes (Shao et al., 2024; Lu et al., 2022; Chen et al., 2024a; Man et al., 2025). ScienceQA (Lu et al., 2022) aimed to enhance model explainability by providing rationales for questions, and subsequently, M3CoT (Chen et al., 2024a) pointed out that ScienceQA was limited to a specific domain and single-step reasoning. M3CoT was thus designed to necessarily require multi-step reasoning across multiple domains, increasing the complexity of inference. These studies marked significant progress in that they explicitly attempted to show the reasoning process through multimodal CoT. However, a limitation shared by both studies is the lack of an explicit link between each reasoning step and specific visual evidence within the image. Grounding the reasoning process in actual visual evidence is essential to verify that the model's logic relies on true image content rather than merely producing plausible text, which fosters trust in the model.

To address this problem of the lack of visual evidence, Visual CoT (Shao et al., 2024) attempted to guide the model's focus by adding a single core region bounding box, which is necessary for the question and answer, as CoT information. However, Visual CoT primarily focused on a single region related to the final answer and was biased toward only one type of evidence, either object or text. This still limits its ability to train complex reasoning skills that require the comprehensive referencing of multiple objects and text evidence during a multi-step reasoning process (Shao et al., 2024; Lu et al., 2022; Chen et al., 2024a). Consequently, previous studies left limitations including (1) difficulty in achieving scalability due to manual construction, (2) the disconnection between the reasoning process and visual evidence, and (3) insufficient utilization of multiple object and text evidence. To simultaneously solve these issues, this study proposes the

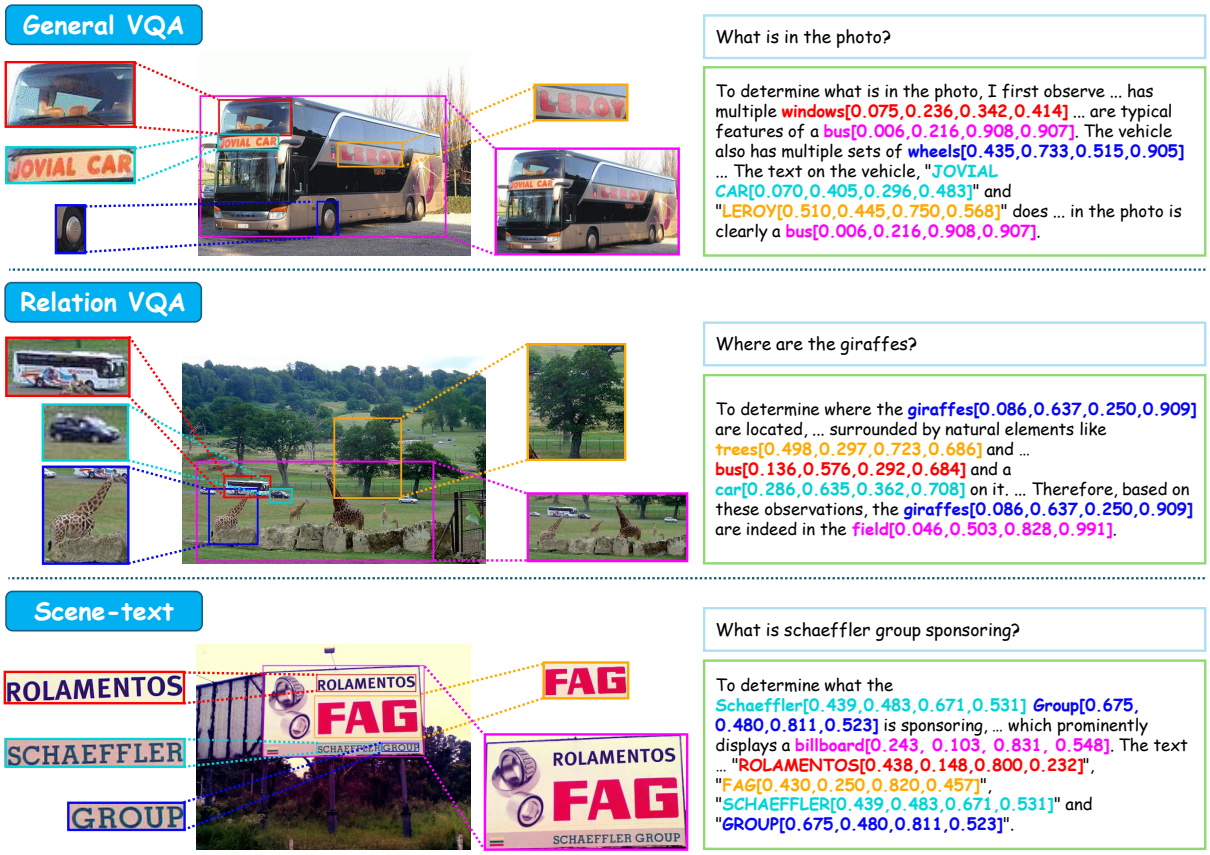


Figure 1: Examples of the Proposed VG-CoT Dataset across Three Task Types.

VG-CoT dataset, which explicitly aligns the location information of multiple objects and text with each reasoning step through an automated three-stage pipeline, thereby laying the groundwork for complex and trustworthy rationale-based visual reasoning.

2.2. Evaluation of Reasoning in LVLMs

Existing benchmarks for evaluating the reasoning capabilities of LVLMs have traditionally been measured solely by the correctness of the model's final answer (Antol et al., 2015; Hudson and Manning, 2019). This approach has a limitation: even if a model provides the correct answer, it cannot distinguish whether the result is due to rational reasoning based on correct visual evidence or merely due to statistical bias or hallucination (Guan et al., 2024; Agarwal et al., 2020; Shah et al., 2019). To overcome the limitations of this simple accuracy-based evaluation, new attempts have emerged to measure various capabilities and the reliability of models. MME (Fu et al., 2024) sought to comprehensively evaluate LVLMs across 14 fine-grained tasks by dividing the model's capabilities into perception and cognition. However, since it primarily focused on the accuracy of the final answer, it had the limitation of not evaluating the intermediate reasoning process that the model used to derive the answer or the visual evidence utilized during that process (Fu et al., 2024; Qiu et al., 2024).

In an effort to evaluate the reasoning process itself, CURE (Chen et al., 2024b) was proposed to measure the consistency of CoT reasoning. It introduced a metric that provides high-level reasoning problems and step-by-step sub-questions to solve them, evaluating whether the model consistently answers not only the final correct answer but also the intermediate-step questions. While this was a significant advancement in evaluating the logical flow of the reasoning process, it failed to directly assess whether the model correctly referenced and utilized specific visual evidence within the image to solve the step-by-step sub-questions. Consequently, the current evaluation system still has a gap in comprehensively verifying the model's fine-grained reasoning ability in terms of 'the logicality of the reasoning process' and 'the utilization of visual evidence' (Jing et al., 2024; Qiu et al., 2024; Guan et al., 2024). Therefore, to fill this gap, this study proposes a new benchmark that comprehensively measures (1) Rationale Quality (visual evidence utilization, logicality, completeness), (2) Answer Accuracy, and (3) Reasoning-Answer Alignment to analyze the trustworthy reasoning capabilities of

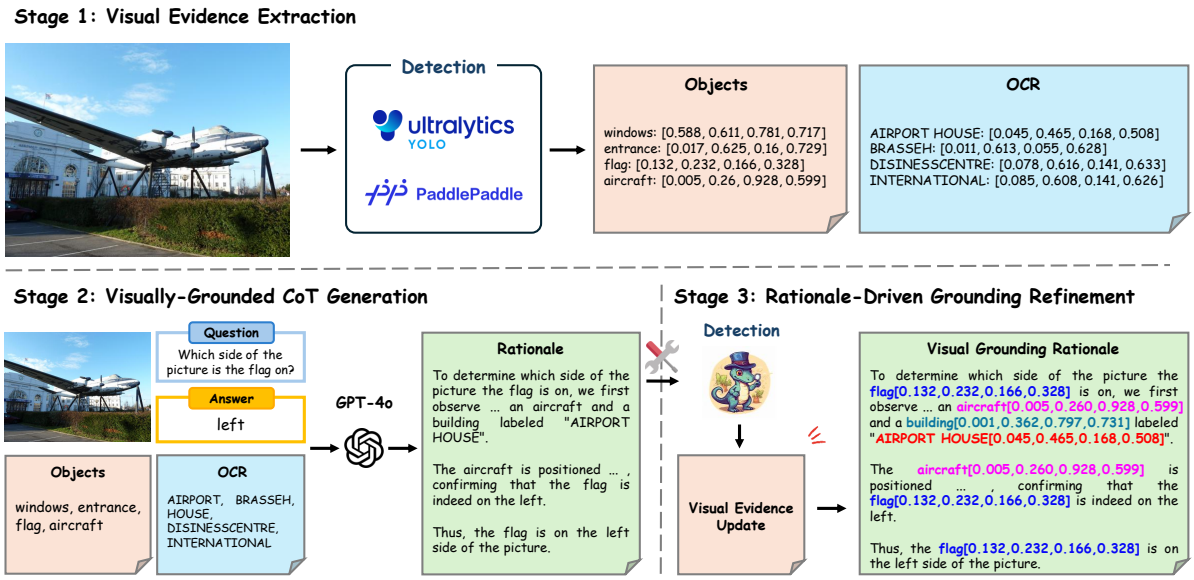


Figure 2: Overview of the Automated Pipeline for Generating Grounded CoT Data.

LVLMs.

3. VG-CoT Dataset

To overcome the limitations of existing datasets, this study introduces the Visual Grounding Chain-of-Thought (VG-CoT) dataset. The construction of VG-CoT is driven by three core objectives: (1) achieving scalability through a fully automated pipeline, (2) securing trustworthy reasoning by explicitly linking each logical step to precise visual evidence (i.e., spatial coordinates), and (3) encompassing diverse reasoning scenarios across multiple domains. As shown in Figure 1, each data sample systematically integrates an image, a question, an answer, and a step-by-step reasoning process densely grounded in multiple object and text locations.

3.1. Data Generation Pipeline

As illustrated in Figure 2, we develop a fully automated three-stage pipeline to construct the VG-CoT dataset without relying on expensive manual annotation. We source raw data from three distinct tasks to comprehensively cover the complex reasoning scenarios that LVLMs may encounter: Scene-text VQA (TextVQA) (Singh et al., 2019) for text-intensive reasoning, General VQA (Visual7W) (Zhu et al., 2016) for object recognition, and Relation VQA (GQA) (Hudson and Manning, 2019) for complex inter-object relational reasoning.

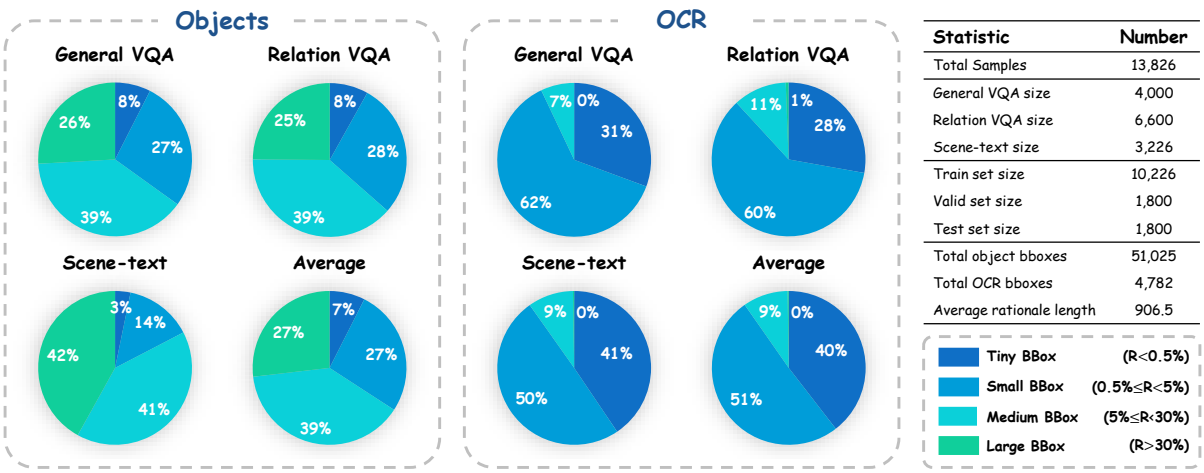
Stage 1: Multi-Granularity Visual Evidence Extraction. The first stage extracts comprehensive visual evidence from the source images to serve as the foundational grounding data. To capture diverse visual elements, we apply specialized extractors for each modality:

- **Object-level Evidence:** The state-of-the-art YOLO model (Redmon et al., 2016) detects the bounding boxes of key objects.
- **Text-level Evidence:** PaddleOCR (Cui et al., 2025) extracts the location and textual content within the image, robustly handling complex environments.

Additionally, for the GQA dataset, the rich scene graph information provided by the dataset itself is leveraged as supplementary initial evidence.

Stage 2: Visually-Grounded CoT Generation. The second stage bridges the gap between visual perception and logical reasoning. We provide GPT-4o, which possesses powerful multimodal capabilities, with the original image, the question, the ground-truth answer, and the visual evidence extracted in Stage 1. The model is explicitly instructed to synthesize this information and generate a step-by-step reasoning path. Crucially, the prompt enforces evidence-based reasoning by requiring the model to directly cite the spatial coordinates (visual evidence) obtained in Stage 1 for every logical deduction it makes toward the final answer.

The specific instructions provided to GPT-4o for this generation task (detailed in Table 1) ensure that the model utilizes the visual evidence fully, moving beyond simple text generation. Crucially, the prompt encourages the model to not only outline



Statistic	Number
Total Samples	13,826
General VQA size	4,000
Relation VQA size	6,600
Scene-text size	3,226
Train set size	10,226
Valid set size	1,800
Test set size	1,800
Total object bboxes	51,025
Total OCR bboxes	4,782
Average rationale length	906.5

Figure 3: Detailed Statistics of the VG-CoT Dataset: Bounding Box Size Distribution (R: Relative Area Ratio) for Objects and OCR, and Key Metrics.

Prompt Template

```
You are an AI assistant that gen-
erates step-by-step reasoning ex-
planations (rationales) for visual
question answering.

{object_info}{ocr_info}

Question: {question}
Answer: {answer}

IMPORTANT GUIDELINES:
1. To solve the question, observe
the image and write the reasoning
process step-by-step.
2. Explicitly connect your rea-
soning process to the visual cues
observed in the image.
3. Include reasoning about why
certain things are NOT the answer
or why other options don't apply.
4. Explain what makes the answer
correct by comparing or contrast-
ing with what is NOT true.
...

Rationale:
```

Table 1: Prompt Template for Generating step-by-step Rationales for VQA.

the path to the correct answer but also to analyze why certain elements are incorrect. This requires the model to demonstrate the plausibility of its answer through comparison and contrast against visual cues, resulting in a highly reliable and faithful VG-CoT.

Stage 3: Rationale-Driven Grounding Refinement. The final stage refines and extends the visual evidence based on the generated reasoning process. While the initial detection in Stage 1 extracts general objects and OCR text, the CoT generated in Stage 2 often references more subtle or context-dependent objects. To align these elements, we employ the following refinement process: First, key object nouns are extracted from the CoT sentences using natural language processing tools (e.g., nltk[1]). These nouns then serve as queries for the open-set detector Grounding DINO (Liu et al., 2025), enabling precise textual grounding of specific objects mentioned in the rationale. Finally, this refined visual evidence is explicitly linked to each reasoning step, ensuring clear alignment between the text and actual image coordinates.

Ultimately, this fully automated pipeline requires no human intervention, achieving high scalability and cost-efficiency while ensuring the consistent quality of the VG-CoT dataset by grounding each reasoning step in visual evidence across diverse scenarios.

3.2. Data Analysis

To comprehensively understand the characteristics of the VG-CoT dataset, this section provides a detailed analysis of its statistics, domain distribution, and visual evidence utilization patterns. By dissecting these aspects, we demonstrate that the dataset possesses rich, multi-dimensional features essential for training and evaluating the multifaceted reasoning capabilities of LVLMs. Figure 3 summarizes

[1] https://www.nltk.org/

the key statistics of the dataset, which comprises a total of 13,826 high-quality samples.

Dataset Scale and Task Distribution. The dataset spans three complementary tasks to cover diverse reasoning abilities, comprising General VQA (4,000 samples) for object understanding, Relation VQA (6,600 samples) for complex inter-object relationships, and Scene-text VQA (3,226 samples) for text comprehension. The total 13,826 samples are partitioned into training (10,226), validation (1,800), and test (1,800) splits at a 7:1.5:1.5 ratio.

Bounding Box Size Distribution. A distinctive feature of the VG-CoT dataset is the stark contrast in bounding box sizes across modalities. As illustrated in Figure 3, while object bounding boxes are predominantly medium or large, text bounding boxes are overwhelmingly tiny or small. This distribution creates a challenging and realistic learning environment, requiring models to precisely identify not only prominent objects but also fine-grained textual evidence.

Multi-Step Grounding Complexity. Unlike previous datasets limited to single-evidence referencing, the multi-step reasoning process in VG-CoT is explicitly linked to a vast amount of visual evidence. Specifically, the dataset incorporates 51,025 total object bounding boxes and 4,782 total OCR bounding boxes. This dense visual grounding is coupled with highly detailed logical steps, reflected by an average rationale length of 906.5. Such complexity provides an environment for training and evaluating the fine-grained, visually-grounded reasoning capabilities of LVLMs.

3.3. Benchmark Construction

To overcome the limitations of existing accuracy-centric evaluations, we establish a comprehensive benchmark designed to systematically assess the reasoning process of LVLMs. This benchmark is structured around three core dimensions.

- **Rationale Quality (RQ).** We utilize a large language model (GPT-4o) as an automated evaluator to assess the quality of the reasoning process on a 1 to 5 scale. This dimension is comprehensively measured through four sub-metrics. **Visual Evidence** evaluates the degree to which the model leverages object and text location information during reasoning. **Coherence** assesses the logical flow between consecutive reasoning steps. **Completeness** measures whether all necessary logical deductions are present. Finally, the **Overall** score represents the combined average of these three aspects.
- **Answer Accuracy (AA).** This metric determines whether the model's final prediction matches the ground truth. It represents the fundamental correctness of the generated output.
- **Reasoning-Answer Alignment (RAA).** Going beyond independent metric evaluations, RAA measures the correlation between the intermediate reasoning quality and the final answer correctness. This allows us to verify how faithfully the reasoning process supports the final prediction. We define two detailed sub-metrics under this dimension.

Consistency Score. This sub-metric calculates the proportion of samples where the rationale quality logically aligns with the answer correctness. A match is defined as either providing excellent reasoning ($RQ \geq 4$) for a correct answer ($AA = 1$) or poor reasoning ($RQ \leq 3$) for an incorrect answer ($AA = 0$).

$$\text{Consistency Score} = \frac{|\mathcal{S}_{\text{consistent}}|}{N} \tag{1}$$

Where:

- $\mathcal{S}_{\text{consistent}} = \{i : (RQ_i \geq 4 \wedge AA_i = 1) \vee (RQ_i \leq 3 \wedge AA_i = 0)\}$
- N is the total number of evaluated samples.
- i is the index of an individual data sample.
- $AA_i \in \{0, 1\}$ denotes the binary correctness of the final answer.

Faithful Score. This sub-metric evaluates the reliability of the model's correct predictions by measuring the proportion of correct answers strictly among the samples that exhibit excellent reasoning ($RQ \geq 4$).

$$\text{Faithful Score} = \frac{|i : RQ_i \geq 4 \wedge AA_i = 1|}{|i : RQ_i \geq 4|} \tag{2}$$

4. Experiment

In this section, we conduct experiments to validate the effectiveness of the proposed VG-CoT dataset and demonstrate the utility of our new benchmark.

4.1. Experimental Setup

Evaluated LVLMs. To validate the effectiveness of our proposed dataset, we perform fine-tuning and evaluation on four representative LVLMs: LLaVA-1.5 (7B and 13B) (Liu et al., 2024), Qwen2-VL-7B-Instruct (Wang et al., 2024), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025).

Model	Rationale Quality				Answer Accuracy	Reasoning-Answer Alignment	
	Visual Evidence	Coherence	Completeness	Overall		Consistency	Faithful
LLaVA-1.5-7B	67.2	77.8	71.6	72.2	48.7	60.9	58.9
+ VG-CoT	**79.4**	**83.9**	**87.0**	**83.4**	**62.5**	**64.1**	**64.3**
LLaVA-1.5-13B	71.3	83.1	74.3	76.2	58.2	**66.9**	**69.7**
+ VG-CoT	**79.8**	**94.3**	**87.7**	**87.3**	**63.2**	65.6	66.2
Qwen2-VL-7B-Instruct	79.1	90.8	84.1	84.7	64.9	66.7	69.1
+ VG-CoT	**81.2**	**94.8**	**88.9**	**88.3**	**72.3**	**72.9**	**75.0**
Qwen2.5-VL-7B-Instruct	80.9	94.9	88.3	88.0	68.5	68.5	70.0
+ VG-CoT	**82.9**	**95.0**	**90.6**	**89.5**	**73.6**	**72.6**	**75.7**

Table 2: Performance Comparison on the Proposed VG-CoT Benchmark.

Model	General VQA		Relation VQA		Scene-text	
	Rationale	Answer	Rationale	Answer	Rationale	Answer
LLaVA-1.5-7B	71.8	51.8	69.8	53.4	63.8	38.0
+ VG-CoT	**86.4**	**69.2**	**84.1**	**64.3**	**86.4**	**43.0**
LLaVA-1.5-13B	76.4	74.0	73.6	59.3	67.7	40.8
+ VG-CoT	**87.9**	**79.8**	**85.1**	**64.4**	**86.2**	**44.8**
Qwen2-VL-7B-Instruct	82.0	70.4	81.9	60.5	84.9	66.6
+ VG-CoT	**89.0**	**81.8**	**86.0**	**65.6**	**88.3**	**73.6**
Qwen2.5-VL-7B-Instruct	88.5	79.4	84.7	57.8	88.2	74.2
+ VG-CoT	**89.9**	**84.4**	**88.2**	**66.4**	**90.6**	**74.8**

Table 3: Task-Specific LVLMs Performance Comparison: Effect of VG-CoT Fine-Tuning.

Benchmarks. To comprehensively evaluate the validity of the models' reasoning capabilities, we utilize our proposed benchmark rather than relying solely on traditional accuracy metrics. The models are evaluated across three core dimensions. First, Rationale Quality (RQ) assesses visual evidence utilization, logical coherence, and reasoning completeness. Second, Answer Accuracy (AA) represents the overall correctness of the model's output. Finally, Reasoning-Answer Alignment (RAA) measures the consistency and faithfulness between the generated rationale and the final prediction.

Implementation Details. All comparative models are fine-tuned using the VG-CoT dataset. During the training phase, the models take an image and a question as input, and are optimized to generate a step-by-step reasoning process explicitly grounded in visual evidence, followed by the final answer. All experiments are conducted in an environment equipped with two NVIDIA H100 (80GB) GPUs. To ensure computational efficiency, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for Parameter-Efficient Fine-Tuning (PEFT). The training process is executed for a total of 3 epochs, utilizing a learning rate of 2e-4 and a batch size of 32.

4.2. Overall Performance on VG-CoT Benchmark

Table 2 presents the performance analysis of four representative LVLMs evaluated on the proposed VG-CoT benchmark. Overall, fine-tuning with the VG-CoT dataset yielded improvements across most dimensions. All evaluated models exhibited consistent gains in both Rationale Quality (RQ) and Answer Accuracy (AA). For instance, the overall RQ of LLaVA-1.5-7B increased from 72.2 to 83.4, while its AA rose from 48.7 to 62.5. This confirms that explicitly linking the reasoning process to visual evidence effectively enhances trustworthy and logical reasoning capabilities.

Despite these overall gains, specific alignment anomalies and grounding bottlenecks were observed. LLaVA-1.5-13B showed a slight decrease in Consistency from 66.9 to 65.6, implying an occasional disconnect between generated rationales and final answers. Furthermore, the Visual Evidence score remained consistently lower than other

	LLaVA-1.5-7B		LLaVA-1.5-13B	
	mAP@0.5	mAP@0.75	mAP@0.5	mAP@0.75
General VQA	46.65	29.07	52.2	36.4
Relation VQA	50.16	29.84	54.9	35.7
Scene-text	30.93	21.64	32.4	26.3
Average	44.5	27.64	48.7	33.6

Table 4: Accuracy of Predicted Visual Evidence for VG-CoT Fine-tuned Models.

RQ components across all models. This indicates that precise spatial grounding is intrinsically more challenging than logical text generation, a trend further supported by the results in Table 4.

4.3. Task-Specific Fine-Tuning Effects and Validity

Table 3 shows the effect of VG-CoT fine-tuning across three distinct reasoning tasks. The results demonstrate that the proposed dataset consistently enhances both Rationale Quality and Answer Accuracy across all evaluated domains, enabling all models to exhibit significant improvements in General and Relation VQA tasks. Most notably, LLaVA-1.5-7B demonstrated remarkable gains of over 10 points in both reasoning and correctness, with its Answer Accuracy jumping from 51.8 to 69.2 in General VQA and from 53.4 to 64.3 in Relation VQA. This confirms that VG-CoT effectively trains models to reason logically about spatial and relational contexts.

The benefits of VG-CoT extend to text-heavy reasoning and already highly capable models. In the challenging Scene-text task, the Rationale score of LLaVA-1.5-7B surged from 63.8 to 86.4. Furthermore, advanced instruction-tuned models like Qwen2.5-VL-7B-Instruct also saw substantial boosts, particularly in General VQA Answer Accuracy, which increased from 79.4 to 84.4. This task-agnostic and consistent improvement comprehensively validates the broad applicability of the VG-CoT dataset.

4.4. Analysis of Visual Evidence Prediction Accuracy

Table 4 details the visual evidence prediction accuracy of the fine-tuned LLaVA models using mean Average Precision (mAP) for bounding box localization against pseudo-label ground truths. As a standard metric for object detection, mAP evaluates the accuracy of predicted bounding boxes by measuring their overlap with ground truth regions at specific Intersection over Union (IoU) thresholds. The results indicate that increased model capacity enhances grounding ability, with LLaVA-1.5-13B outperforming the 7B variant in average mAP@0.5 by achieving 48.7 compared to 44.5. Across tasks, models achieved their highest scores in Relation VQA, reaching 54.9 mAP@0.5 for the 13B model, but struggled significantly in the Scene-text task with a score of 32.4. This substantial gap quantitatively demonstrates that while models can reliably ground larger relational objects, precisely locating small and dense text regions remains intrinsically challenging.

Furthermore, the data reveals a steep performance drop when applying a stricter IoU threshold. For instance, the average precision of LLaVA-1.5-13B drops from 48.7 at mAP@0.5 to 33.6 at mAP@0.75. This highlights that although the VG-CoT fine-tuned models successfully learn to approximate the general location of relevant visual evidence, achieving highly precise spatial grounding is still a significant bottleneck that requires further exploration.

5. Conclusion

This study aimed to address the limitations of existing datasets and benchmarks to enhance the trustworthy rationale-based reasoning capabilities of LVLMs. To this end, we proposed the VG-CoT dataset and an automated three-stage pipeline for its construction, enabling the scalable and reliable generation of visual evidence-based CoT data. Additionally, we established a new benchmark that comprehensively measures Rationale Quality, Answer Accuracy, and Reasoning-Answer Alignment to deeply evaluate the model's reasoning process. Experimental results confirmed that VG-CoT consistently improved reasoning quality and answer accuracy, and also showed improvements in consistency for most models. Furthermore, an analysis of the model's prediction accuracy for visual evidence suggested the potential for improving grounding capability along with future challenges. This study provides a core dataset and evaluation methodology for trustworthy LVLMs reasoning and is expected to contribute to the advancement of related research in the future.

Limitations

Although this study contributes to enhancing visual evidence-based reasoning in LVLMs, it has certain limitations that suggest directions for future work.

First, as the proposed VG-CoT framework is a fully automated three-stage pipeline designed for scalability, its performance is inherently tied to the integration of its underlying foundational models, such as YOLO, PaddleOCR, Grounding DINO and GPT-4o. While this cohesive approach enables cost-efficient dataset construction, the final quality remains subject to the precision of these mod-

els. Furthermore, while we focused on validating the overall effectiveness and scalability of the integrated system, a more granular analysis to disentangle the isolated impact of each stage would provide deeper insights into the grounding mechanism.

Second, while the VG-CoT dataset covers a wide range of general and relational reasoning tasks, its applicability to highly specialized professional domains, such as medical diagnostics or engineering diagram comprehension, has yet to be fully explored. Expanding the pipeline to these specific fields remains a promising direction for future research.

Ethics Statement

This study was conducted using publicly available datasets (GQA, Visual7W, TextVQA) that contain no personally identifiable or sensitive information. The proposed automated pipeline does not involve human annotation or data collection from individuals, thereby minimizing ethical risks. All experiments and analyses were performed in compliance with the ethical standards of dataset usage and research transparency guidelines.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

6. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024a. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024b. Measuring and improving chain-of-thought reasoning in vision-language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 192–210, Mexico City, Mexico. Association for Computational Linguistics.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. 2025. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 18.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang

Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, Miami, Florida, USA. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2024. TopViewRS: Vision-language models as top-view spatial reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1786–1807, Miami, Florida, USA. Association for Computational Linguistics.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564. Curran Associates, Inc.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2025. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Computer Vision – ECCV 2024*, pages 38–55, Cham. Springer Nature Switzerland.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. 2025. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14268–14280.

Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. 2024. Synthesize Diagnose and Optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13279–13288.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. 2024. VALOR-EVAL: Holistic coverage and faithfulness evaluation of large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1783–1805, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual CoT: Advancing multimodal language models with a comprehensive

dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 8612–8642. Curran Associates, Inc.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, et al. 2025. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. *arXiv preprint arXiv:2509.25373*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR.