

# Weaving Context Across Images: Improving Vision-Language Models through Focus-Centric Visual Chains

Juntian Zhang<sup>1</sup>, Chuanqi Cheng<sup>1</sup>, Yuhan Liu<sup>1\*</sup>, Wei Liu<sup>2</sup>, Jian Luan<sup>2</sup>, Rui Yan<sup>1,3,4\*</sup>

<sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China <sup>2</sup>MiLM Plus, Xiaomi Inc.


<sup>3</sup>Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MoE

<sup>4</sup>School of Artificial Intelligence, Wuhan University

{zhangjuntian, chengchuanqi, yuhan.liu, ruiyan}@ruc.edu.cn

{liuwei40, luanjian}@xiaomi.com

## Abstract

Vision-language models (VLMs) achieve remarkable success in single-image tasks. However, real-world scenarios often involve intricate multi-image inputs, leading to a notable performance decline as models struggle to disentangle critical information scattered across complex visual features. In this work, we propose *Focus-Centric Visual Chain*, a novel paradigm that enhances VLMs' perception, comprehension, and reasoning abilities in multi-image scenarios. To facilitate this paradigm, we propose *Focus-Centric Data Synthesis*, a scalable bottom-up approach for synthesizing high-quality data with elaborate reasoning paths. Through this approach, we construct **VISC-150K**, a large-scale dataset with reasoning data in the form of *Focus-Centric Visual Chain*, specifically designed for multi-image tasks. Experimental results on seven multi-image benchmarks demonstrate that our method achieves average performance gains of **3.16%** and **2.24%** across two distinct model architectures, without compromising the general vision-language capabilities. Our study represents a significant step toward more robust and capable vision-language systems that can handle complex visual scenarios:  **VISC**.

## 1 Introduction

The rapid advancement of VLMs has revolutionized traditional visual tasks with single-image input, achieving human-level performance in various applications (Daniali and Kim, 2023). However, real-world scenarios frequently involve more complex visual input, such as multiple images, where current VLMs show significant performance degradation (Zhao et al., 2024). The challenges stem from two complementary traits of multi-image tasks: (1) **Cross-image correlations**: Images are often diversely related, requiring a holistic understanding of their contextual relationships. (2) **Vi-**

\* Corresponding authors.

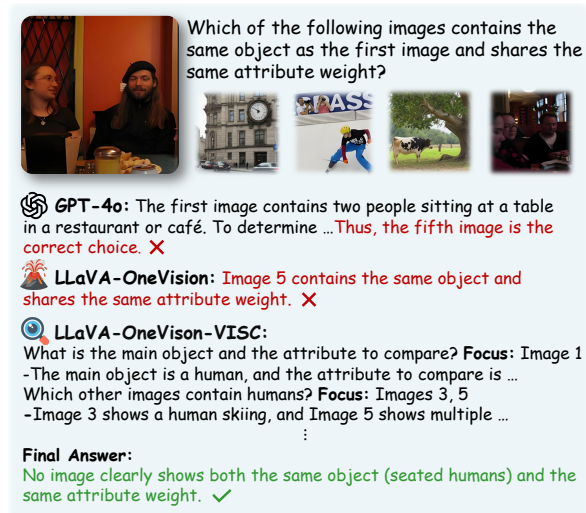


Figure 1: A multi-image QA example: Using *Focus-Centric Visual Chain*, our model LLaVA-OneVision-VISC successfully answers a question that both GPT-4o and LLaVA-OneVision fail to solve correctly.

**sual discontinuity**: The fragmentation of information between images makes it challenging to accurately grasp cross-image relationships.

To address these challenges, we propose *Focus-Centric Visual Chain*, a multi-image reasoning paradigm that progressively focuses on key information. In this process, VLMs decompose a complex task into a sequence of sub-processes, each involving the resolution of a sub-question that focuses on localized visual input. As illustrated in Figure 1, this process allows models to iteratively aggregate the necessary visual evidence guided by the sub-questions, ultimately leading to the resolution of the complex task.

To implement this paradigm, high-quality reasoning data is indispensable. However, despite extensive research on reasoning tasks, reasoning data for multi-image scenarios remains scarce. While there exist approaches leveraging multimodal models to directly generate reasoning chains (Zhang

et al., 2024c) or distilling data from more muscular models, such methods present two primary limitations: (1) **Insufficient reliability**, even state-of-the-art closed-source models (e.g., GPT-4o) demonstrate inconsistent performance on multi-image tasks (Wang et al., 2024a); and (2) **Prohibitive costs**, the substantial cost of closed-source models severely constrains scalability. Hence, we propose the *Focus-Centric Data Synthesis* framework, an efficient approach for generating reasoning data following a bottom-up strategy. In contrast to our reasoning paradigm, which decomposes complex tasks into simpler sub-tasks, the framework centers on progressive information aggregation. At each stage, it expands the existing information set to formulate reliable reasoning paths and corresponding questions from complex visual inputs.

The *Focus-Centric Data Synthesis* framework comprises four systematically designed modules: (1) *Feature Extraction* constructs comprehensive textual profiles for each image, which serve as nodes in subsequent process; (2) *Pair Connection* identifies relevant image pairs through object-oriented and event-oriented detection, forming edges between distinct nodes; (3) *Relevance Annotation* categorizes detected connections into three predefined types (Temporal, Spatial, and Semantic) and detail them; (4) *Question Generation* produces logically chained sub-questions based on the established inter-image network, ultimately synthesizing the final composite question and reasoning path. The framework’s bottom-up design ensures data quality while maintaining computational efficiency through exclusive use of open-source models.

Leveraging this framework, we construct **VISC-150K** composed of 150K high-quality multi-image reasoning samples. Extensive experiments across seven multi-image benchmarks demonstrate the effectiveness of our approach. When integrated with different base models, VISC-150K consistently brings performance improvements across all challenging benchmarks, with average accuracy increased by 3.16% and 2.24%, respectively, achieving new state-of-the-art on four out of the seven.

In summary, our contributions are three-fold:

- We introduce the *Focus-Centric Visual Chain* paradigm to solve complex multi-image tasks through question decomposition and stepwise reasoning.
- We propose *Focus-Centric Data Synthesis*, a framework tackling data scarcity by synthesiz-

ing reliable, cost-effective, and reproducible reasoning data via open-source models.

- We release VISC-150K, a multi-image reasoning dataset containing 150K data with Focus-Centric Visual Chains. Our dataset delivers consistent performance gains across diverse model architectures on seven challenging multi-image benchmarks, as validated by comprehensive evaluations.

## 2 Related Work

### 2.1 Vision-Language Models

Vision-Language Models (VLMs) integrate visual and textual processing through multimodal architectures, empowering various types of tasks (Liu et al., 2024c,d). Closed-source models such as GPT-4o (OpenAI, 2024a) and Gemini-1.5-Pro (Gemini Team, 2024) demonstrate state-of-the-art performance. While open-source VLMs can be categorized into two types based on their architecture. The first type employs a unified architecture for both visual and textual modality. For example, Flamingo (Alayrac et al., 2022) incorporates visual information into textual inputs through blocks based on cross-attention. The KOSMOS series (Huang et al., 2023; Peng et al., 2023) uses the same embedding module to encode text and visual information.

The second type aligns the two modalities by utilizing a connector module to project visual inputs into the textual space. BLIP-2 (Li et al., 2023) connects the visual encoder and the language model with Q-former. InstructBLIP (Dai et al., 2023) further proposes an innovative instruction integration to achieve better modality alignment. The LLaVA series (Liu et al., 2023) adopts a more concise design, using MLP as the mapping layer between modalities. This architecture was widely adopted by recent VLMs, including Mantis (Jiang et al., 2024), LLaVA-OneVision (Li et al., 2025), and InternVL2 (Chen et al., 2024c). Additionally, Qwen2-VL (Wang et al., 2024b) adopts visual position encoding and Naive Dynamic Resolution.

Despite advancements, evaluations reveal two critical limitations of previous works: (1) VLMs struggle with complex multi-image tasks (Zhao et al., 2024) and (2) existing training paradigms offer diminishing returns for multi-image tasks (Campbell et al., 2024). To overcome these limitations, we synthesize a dataset that continuously improves VLMs’ performance on various

multi-image benchmarks, thus providing a new solution for complex multi-image scenarios.

## 2.2 Reasoning of LLMs and VLMs

Reasoning capabilities in Large Language Models (LLMs) have evolved through innovative prompting strategies and knowledge integration. Chain-of-Thought (CoT) (Wei et al., 2022) pioneered reasoning in LLM, later enhanced by Tree-of-Thought (Yao et al., 2023) and Graph-of-Thought (Besta et al., 2024). Self-consistency (Wang et al., 2023) improves robustness by aggregating multiple reasoning paths via voting strategy. Recent advances like OpenAI’s o1 series (OpenAI, 2024b) employ reinforcement learning to foster more complex reasoning.

Beyond focusing on single-modal, multimodal reasoning (Wang et al., 2024c) extensions leverage data-centric approaches (Gao et al., 2023; Zhang et al., 2024b; Shi et al., 2024; Cheng et al., 2024; Xu et al., 2024), knowledge graph integration (Zhang et al., 2024a; Lee et al., 2024), and tree search (Yao et al., 2024) for transferring reasoning capabilities to VLMs. However, for scenarios with multi-image inputs, how to enable VLMs to integrate complex visual information and perform slow thinking remains an issue that has not been fully explored. To address this issue, we design a multi-step reasoning paradigm, decomposing the initial question into a sequence of sub-questions and focusing on a subset of input images in each step. The effectiveness of this paradigm is verified by experiments in § 4.

## 3 Methodology

The methodology is detailed in two parts. First, we formulate the *Focus-Centric Visual Chain*, which enhances the ability of VLMs to handle complex multi-image inputs through multi-step reasoning (§ 3.1). Then, we present the *Focus-Centric Data Synthesis* framework, a bottom-up approach for synthesizing multi-image reasoning data with precise control over the reasoning process (§ 3.2).

### 3.1 Focus-Centric Visual Chain

We present *Focus-Centric Visual Chain*, a reasoning paradigm specifically designed for multi-image input scenarios. Given an image collection  $\mathcal{G} = \{I_k | k = 1, 2, \dots, K\}$  and an initial question  $Q$ , the model  $\mathcal{M}$  progressively constructs a reasoning chain  $\mathcal{R}$  via multi-step reasoning. At each

step, the model focuses on a visual evidence subset  $G \subseteq \mathcal{G}$  through a dynamic selection mechanism.

Specifically, at the  $i$ -th reasoning step, the model  $\mathcal{M}$  generates an intermediate sub-question  $q_i$  and identifies its corresponding focus, a minimized visual information subset  $G_i$ , through:

$$q_i, G_i = \mathcal{M}(Q, \mathcal{G}, \mathcal{H}_{<i}), \quad (1)$$

where  $\mathcal{H}_{<i} = \{q_1, q_2, \dots, q_{i-1}\}$  represents the sub-question history of previous steps, with  $\mathcal{H}_{<1} = \emptyset$  for initialization. Subsequently, the model derives an intermediate answer  $a_i$  through joint analysis of  $q_i$  and  $G_i$ . According to the existing answer set  $\mathcal{A}_i = \{a_1, a_2, \dots, a_i\}$ , the model determines whether to extend the reasoning path and outputs a stopping signal  $z_i \in \{True, False\}$ . Only when  $z_i = True$ , does the model synthesize the final answer  $A$  from the current QA collection  $\mathcal{C}_i = \{(q_1, a_1), (q_2, a_2), \dots, (q_i, a_i)\}$  and terminate the reasoning; otherwise, it iteratively refocuses on distinct visual information. Therefore, the overall reasoning process  $\mathcal{R}$  can be represented as an ordered sequence:

$$\mathcal{R} = [(q_i, G_i, a_i, z_i)]_{i=1}^N, \quad (2)$$

where  $N$  denotes the total number of executed reasoning steps.

### 3.2 Focus-Centric Data Synthesis

Aiming to implement the reasoning paradigm above in a data-driven manner, we propose *Focus-Centric Data Synthesis (FCDS)*, a bottom-up evolutionary data synthesis framework. *FCDS* cost-effectively facilitates the large-scale synthesis of certifiable cross-image reasoning data through open-source models. The synthesis process begins with a set of images and incrementally assesses their interrelationships. It then formulates intermediate sub-questions, culminating in a meaningful question that aligns with a coherent reasoning path. Specifically, *FCDS* consists of four interconnected steps: *Feature Extraction*, *Pair connection*, *Relevance Annotation* and *Question Generation*, as illustrated in Figure 2.

**Feature Extraction.** Following a bottom-up manner, the synthesis process initiates with granular feature extraction, constructing a detailed profile for each image  $I \in \mathcal{G}$ , which consists of four core elements: (1) the overall view of  $I$ ; (2) background descriptions; (3) object attributes and (4) object interactions. These image profiles provide a pathway

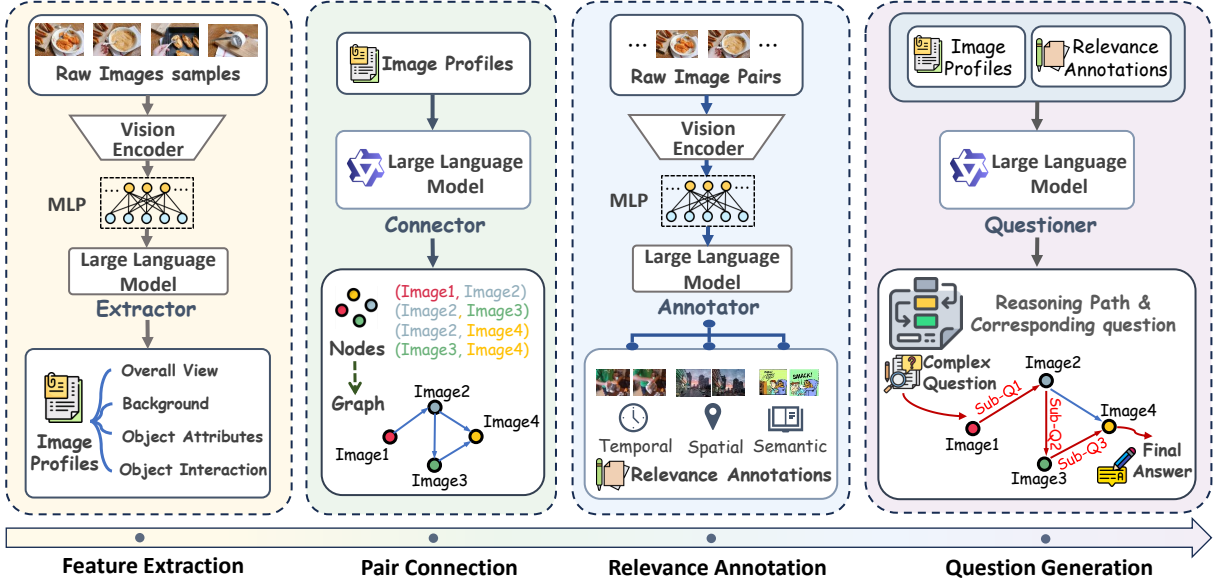


Figure 2: The *Focus-Centric Data Synthesis* framework consists of four stages: **Feature Extraction** processes raw visual input, extracting object-level attributes and interactions into structured image profiles. **Pair Connection** links related image nodes based on their profiles. **Relevance Annotation** identifies and annotates relationships between nodes from temporal, spatial, and semantic perspectives. **Question Generation** utilizes the extracted image profiles and relationship annotations to construct multi-image reasoning paths and corresponding questions.

for capturing object-level features and modeling their relationships. All profiles are generated by our specialized vision-language model **Extractor**, which comprises three fundamental components: a visual encoder  $f_e$  for visual feature encoding, a vision-language connector  $f_c$  for modality alignment, and a large language model  $f_\phi$  for semantic understanding as well as content generation.

Specifically, for each image  $I \in \mathcal{G}$ , the visual encoder  $f_e$  projects  $I$  into a sequence of latent embeddings:

$$X^V = f_e(I) = \langle x_1^V, x_2^V, \dots, x_n^V \rangle, \quad (3)$$

where each visual token  $x_i^V \in \mathbb{R}^{d^V}$  corresponds to an image patch, with  $d^V$  denoting the output dimension of the visual encoder. The number of visual tokens  $n$  depends on the visual encoder<sup>1</sup>. Next, the vision-language connector  $f_c$  performs dimension-aware projection to align visual features with textual semantics:

$$X^T = f_c(X^V) = \langle x_1^T, x_2^T, \dots, x_n^T \rangle, \quad (4)$$

where each  $x_i^T \in \mathbb{R}^{d^T}$  represents an aligned token. Here,  $d^T$  is the dimension of  $f_\phi$ .  $X^T$  is then fed into the large language model  $f_\phi$  to generate the

<sup>1</sup>In some VLMs (e.g., LLaVA-OneVision),  $n$  is a fixed number. While for some VLMs (e.g., Qwen2-VL),  $n$  increases with the resolution of  $I$ .

profile  $p = f_\phi(X^T)$  for  $I$ . Each generated profile is treated as a node in the reasoning path.

**Pair Connection.** Since connecting arbitrary nodes may lack semantic validity, we establish edges only between nodes with potential relevance. We propose two criteria to determine node relatedness: (1) **Object-oriented**, where images share co-occurring objects, and (2) **Event-oriented**, where images depict shared or related events. Given profile collection  $\mathcal{P}$  of image set  $\mathcal{G}$ , a large language model **Connector** is implemented to identify valid pairwise connections:

$$\{(i, j) | i, j \in [0, K], i \neq j\} = \text{Connector}(\mathcal{P}), \quad (5)$$

where each pair  $(i, j)$  indicates a potential connection between image  $I_i$  and  $I_j$ . By identifying correlations between different nodes, we establish the basic structure of the reasoning path.

**Relevance Annotation.** To systematically characterize inter-node relevance, we classify it into three categories: Temporal, Spatial, and Semantic, as detailed as follows:

- **Temporal:** The paired images depict a chronological sequence, with one distinctly preceding the other in temporal succession.
- **Spatial:** Visual elements in paired images exhibit geometric and positional correlations, forming spatial continuity or progression.

- **Semantic:** The paired images exhibit intangible associations containing thematic, logical and causal relationships. Notably, this category accommodates abstract connections beyond direct visual correspondence.

To formalize the relation annotation process, we develop **Annotator**, which also comprises three components: a visual encoder  $f_e$ , a vision-language connector  $f_c$ , and a large language model  $f_\phi$ . Given connected image pair  $s = (I_i, I_j)$ , **Annotator** first performs independent encoding of both images followed by feature concatenation:

$$X_i^V = f_e(I_i) = \langle x_{i,1}^V, x_{i,2}^V, \dots, x_{i,n}^V \rangle, \quad (6)$$

$$X_j^V = f_e(I_j) = \langle x_{j,1}^V, x_{j,2}^V, \dots, x_{j,n}^V \rangle, \quad (7)$$

$$X^V = \text{concat}(X_i^V, X_j^V), \quad (8)$$

where  $X^V$  denotes the concatenated embeddings of the encoded image pair, resulting in a sequence of  $2n$  visual tokens. Subsequently,  $X^V$  is mapped to aligned textual tokens  $X^T$  through the vision-language connector  $f_c$ :

$$X^T = f_c(X^V) = \langle x_{i,1}^T, \dots, x_{i,n}^T, x_{j,1}^T, \dots, x_{j,n}^T \rangle. \quad (9)$$

Finally, the large language model  $f_\phi$  generates the relation  $r = f_\phi(X^T)$  where  $r$  denotes the annotated relation for  $s$ . Notably, multiple relation types may be simultaneously present in a single pair.

**Question Generation.** The reasoning path is constructed by sampling a sequential chain of  $K$  interconnected nodes. For each connected image pair  $s_i$  along this path, we generate targeted sub-questions  $q_i$  based on their annotated relations  $r_i$  and corresponding image profiles. The question generation process is implemented through a specialized large language model **Questioner**:

$$q_i = \text{Questioner}(s_i, r_i, p_{i,1}, p_{i,2}), \quad (10)$$

where  $p_{i,1}$  and  $p_{i,2}$  represent the profiles of the two images in  $s_i$ . Finally, these sub-questions are synthesized into a coherent overarching question  $Q$  through aggregation:

$$Q = \text{Questioner}(\{q_i\}_{i=1}^K). \quad (11)$$

We utilized LLaVA-OneVision-7B <sup>2</sup> (Li et al., 2025) as the base model for the **Extractor** and

<sup>2</sup><https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov>

**Annotator**, while Qwen2.5-7B-Instruct <sup>3</sup> (Team, 2024) served as the base model for the **Connector** and **Questioner**. Through our carefully designed framework, we constructed VISC-150K, a high-quality dataset comprising 150K multi-image reasoning data instances following the *Focus-Centric Visual Chain* paradigm. The image resources are collected from publicly accessible websites and include real-world photographs with diverse scenes and comics. More details about our dataset are introduced in Appendix A.

## 4 Experiments

We first evaluate the effectiveness of our method across diverse multi-image tasks. We then conduct more investigations into our method through multifaceted experimental studies and in-depth analysis.

### 4.1 Experimental Setup

We apply *Focus-Centric Visual Chains* to two pre-trained models, LLaVA-OneVision-7B and Qwen2-VL-7B-Instruct <sup>4</sup> (Wang et al., 2024b), which have been extensively fine-tuned on large-scale multi-image datasets and exhibit robust capability. Both models undergo LoRA (Hu et al., 2022) fine-tuning on VISC-150k for one epoch with a batch size of 8. The learning rate is set to  $1e-5$  with a warmup ratio of 0.05 and as a cosine scheduler. The maximum context length is set to 32,768.

When conducting evaluations, the temperature is set to 0 and the max new tokens is 1,024. For Qwen2-VL-7B, the image resolution is cropped between  $128 \times 28 \times 28$  and  $1280 \times 28 \times 28$  to reduce memory consumption and improve inference speed. More details about experimental settings are reported in Appendix B.

### 4.2 Baselines

For open-source VLMs, our baselines include Qwen2-VL (Wang et al., 2024b), Qwen-VL-Chat (Bai et al., 2023), LLaVA-OneVision (Li et al., 2025), LLaVA-1.6 (Liu et al., 2024b), LLaVA-1.5 (Liu et al., 2024a), InternVL2.5 (Chen et al., 2024b), InternVL2 (Chen et al., 2024c), Mantis-Idetics2 (Jiang et al., 2024), Idetics2 (Laurençon et al., 2024), VILA-1.5 (Lin et al., 2024) and OpenFlamingo-v2 (Awadalla et al., 2023). Among close-source VLMs, we select GPT-4V/GPT-4o (OpenAI, 2024a) as the baseline. Please refer to

<sup>3</sup><https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

<sup>4</sup><https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>

Model	Size	MMIU	MuirBench	MIRB	BLINK	NLVR2	Mantis-Eval	MVBench
GPT-4V/GPT-4o	-	55.70	68.00	53.05	51.14	88.80	62.67	43.50
Qwen-VL-Chat	7B	15.90	20.42	14.38	31.17	58.72	39.17	42.15
LLaVA-1.5	7B	19.20	23.46	28.47	37.13	53.88	31.34	36.00
LLaVA-1.6	7B	22.20	27.42	29.83	39.55	58.88	45.62	40.90
Idefics2	8B	27.80	26.08	33.02	45.18	86.87	48.85	29.68
VILA-1.5	8B	32.45	33.12	36.52	39.30	76.45	51.15	49.40
OpenFlamingo-v2	9B	22.30	23.73	28.80	39.18	36.41	12.44	7.90
Mantis-Idefics2	8B	45.60	44.50	34.82	49.05	89.71	57.14	51.38
InternVL2	8B	42.00	48.70	50.00	50.90	-	65.40	65.80
InternVL2.5	8B	46.70	<b>51.10</b>	52.50	<u>54.80</u>	-	67.70	<b>72.00</b>
LLaVA-OneVision	7B	40.32	41.77	51.18	48.20	89.40	64.20	56.70
+VISC-150K	7B	46.52 <sub>(↑6.20)</sub>	49.62 <sub>(↑7.85)</sub>	53.02 <sub>(↑1.84)</sub>	50.24 <sub>(↑2.04)</sub>	<b>89.88</b> <sub>(↑0.48)</sub>	66.36 <sub>(↑2.16)</sub>	58.23 <sub>(↑1.53)</sub>
Qwen2-VL	7B	50.00	39.12	<u>58.67</u>	53.20	86.42	<b>69.60</b>	67.00
+VISC-150K	7B	<b>52.76</b> <sub>(↑2.76)</sub>	44.50 <sub>(↑5.38)</sub>	<b>60.16</b> <sub>(↑1.49)</sub>	<b>55.34</b> <sub>(↑2.14)</sub>	<u>89.82</u> <sub>(↑3.40)</sub>	<u>69.12</u> <sub>(↓0.48)</sub>	<u>68.01</u> <sub>(↑1.01)</sub>

Table 1: Performance comparison of LLaVA-OneVision and Qwen2-VL based models across seven multi-image benchmarks, with the highest scores being **bolded** and the second highest underlined. Results highlighted in gray indicate experiments using closed-source models, while the remaining results are from open-source models.

Appendix C for more details.

### 4.3 Benchmarks

We evaluate our method on seven comprehensive multi-image benchmarks that span diverse scenarios. The statistics of each benchmark are detailed in Appendix D, with characteristics listed as follows: (1) **MMIU** (Meng et al., 2025) categorizes multi-image relationships into three primary types: semantic, spatial, and temporal. These categories are further subdivided into seven subtypes, covering 52 distinct multi-image understanding tasks.

(2) **MuirBench** (Wang et al., 2024a) covers 12 distinct multi-image understanding tasks and encompassing 10 types of multi-image relationships. (3) **MIRB** (Zhao et al., 2024) includes four evaluation dimensions: perception, visual world knowledge, reasoning, and multi-hop reasoning. Each category consists of tasks requiring comparison and inference across multiple images.

(4) **BLINK** (Fu et al., 2024) incorporates 14 visual perception tasks that humans can quickly solve, covering indoor, outdoor, and natural scenes.

(5) **NLVR2** (Suhr et al., 2019) contains examples of English sentences paired with online photos, focusing on rich linguistic and visual content to support diverse reasoning tasks.

(6) **Mantis-Eval** (Jiang et al., 2024) comprises high-quality multi-image reasoning samples, designed for diverse multi-image skills such as co-reference, reasoning, and comparison.

(7) **MVBench** (Li et al., 2024) consists of 20 challenging video understanding tasks, which cover a wide range of temporal understanding skills in

video scenarios.

### 4.4 Results

The experimental results are presented in Table 1. Both LLaVA-OneVision and Qwen2-VL demonstrate consistent performance improvements across seven benchmarks after fine-tuning with our synthesized dataset VISC-150K, indicating the effectiveness of our method across different VLM architectures.

LLaVA-OneVision achieves benchmark-leading improvements of 6.20% on MMIU and 7.85% on MuirBench, demonstrating breakthrough capabilities. Significantly, our method elevates even the already superior Qwen2-VL model across multiple benchmarks, achieving an average gain of 2.24% over its strong baselines. When combined with LLaVA-OneVision’s 3.16% average improvement, these consistent enhancements across both high-performing and emerging models conclusively validate the universality of our approach.

Our method establishes new state-of-the-art results on four benchmarks: MMIU, MIRB, BLINK, and NLVR2. These improvements can be attributed to three key characteristics: (1) the richness of visual information, (2) the diversity of inter-image relationships, and (3) the complexity of task formulations. These characteristics align well with our method’s enhanced capabilities in visual perception, comprehension, and reasoning.

Moreover, our method demonstrates measurable performance gains on the video benchmark MVBench. Given that a video is essentially a collection of frames, it falls into a multi-image sce-

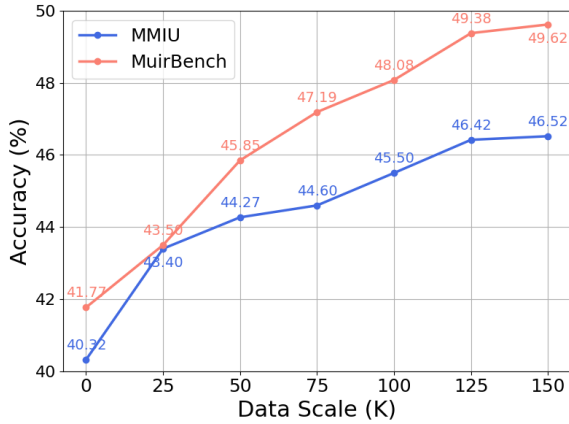


Figure 3: The impact of dataset scale on LLaVA-OneVision’s performance across MMIU and MuirBench benchmarks. As the data scale increases, the model’s accuracy progressively improves.

nario with temporal correlations, where our method maintains efficacy. This result substantiates that our *Focus-Centric Visual Chain* reasoning paradigm operates in a domain-agnostic manner, exhibiting robust applicability as well as transferability.

#### 4.5 Discussions

In this section, we propose five research questions and conduct an in-depth investigation to provide a more comprehensive evaluation of our method.

##### RQ1: How does data size affect performance?

To investigate the impact of the data scale, we create five subsets from VISC-150K through random sampling, which contain 25K, 50K, 75K, 100K, and 125K instances respectively. Each subset is used to fine-tune LLaVA-OneVision using LoRA for one epoch. The fine-tuned models are then evaluated on the MMIU and MuirBench, and the result is visualized in Figure 3.

The performance curves reveal a non-linear relationship: the model achieves rapid improvements when increasing data size from 0 to 125K, followed by a more gradual improvement from 125K to 150K, suggesting diminishing returns but continued learning potential. We attribute the rapid performance leap observed with the 0–25K data scale to a capability activation process, where the model unlocks stronger multi-image potential by learning from reasoning data under the new paradigm. This result suggests that data constructed through *Focus-Centric Data Synthesis* framework can be effectively scaled up, which is essential for further enhancing performance by expanding the data size.

##### RQ2: What is the effect on different sub-tasks?

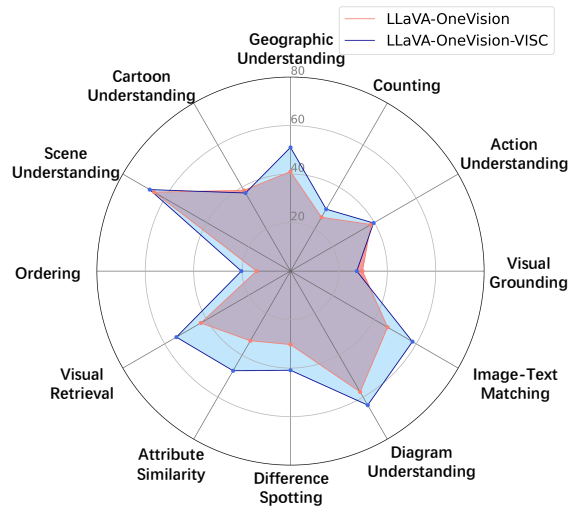


Figure 4: The accuracy comparison of LLaVA-OneVision on 12 MuirBench sub-tasks with and without being fine-tuned on VISC-150K.

To better understand our approach’s strengths, we conducted a detailed analysis across different multi-image tasks in MuirBench. We compare the performance of LLaVA-OneVision before and after being finetuned on VISC-150K, as illustrated in Figure 4. Across all twelve sub-tasks in MuirBench, only four of them do not exhibit substantial improvements. These tasks face dual constraints stemming from (1) intrinsic limitations in vision-language architectures and (2) capacity constraints of foundational language models. This is exemplified by the 3D spatial reasoning required for *Scene Understanding* and the nuanced semantic interpretation essential for *Cartoon Understanding*.

In contrast, the remaining eight tasks show significant improvements. These tasks predominantly involve similarity analysis or comparative reasoning at the image or feature level, such as visual retrieval and attribute similarity. Additionally, some of these sub-tasks (e.g., geographic understanding) involve image types and task categories that are not present in VISC-150K, which emphasize the cross-domain generalization capabilities of our method.

##### RQ3: How does the number of input images impact performance?

To investigate how the number of input images impacts our method, we conducted a detailed analysis of LLaVA-OneVision on the MMIU benchmark. The instances are grouped into different buckets by the number of input images.

As demonstrated in Figure 5, when handling 3-8 images, LLaVA-OneVision-VISC achieves remark-

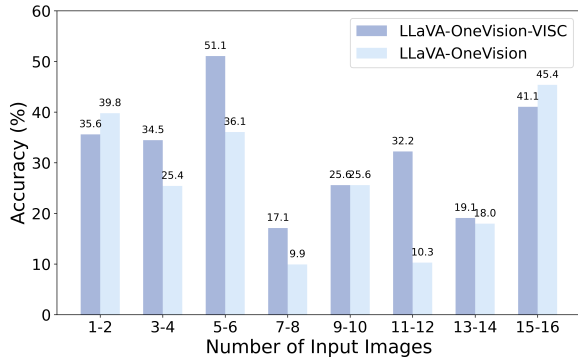


Figure 5: The distribution of task accuracy for LLaVA-OneVision based models across varying numbers of input images, grouped into eight buckets from 1 to 16 images with an interval of 2.

able improvements, suggesting successful identification of cross-image relationships in medium-sized image sets. This capability persists even with larger inputs (11-14 images), where LLaVA-OneVision-VISC maintains superior performance compared to baseline models while avoiding performance degradation from information overload. However, when processing more than 15 images, the performance of LLaVA-OneVision-VISC exhibits slight degradation, which may be attributed to amplified noise levels or interference from irrelevant data patterns in prolonged image sequences.

#### RQ4: Does VISC-150K affect general ability?

Although our method demonstrates remarkable improvements in multi-image tasks, it is crucial to evaluate whether these gains come at the expense of general task performance. To investigate this issue, we adopt Qwen2-VL as the base model and conduct analysis on four benchmarks in other domains. These benchmarks are based on single-image inputs and evaluate the fine-tuned model’s performance from various perspectives, including hallucination, single-image data language capability, domain-specific knowledge, and mathematical reasoning. Specifically: **HallusionBench** (Guan et al., 2024) is designed to assess VLMs’ ability to comprehend and interpret visual data; **MMStar** (Chen et al., 2024a) requires advanced multi-modal capabilities for accurate interpretation; **MMMU** (Yue et al., 2024) focuses on evaluating models’ ability to apply domain-specific knowledge; **MathVista** (Lu et al., 2024) integrates mathematical reasoning with visual tasks. The results are presented in Table 2. The model fine-tuned on VISC-150K maintains comparable or superior performance across all benchmarks, indicating that our

Model	HallusionBench	MMStar	MMMU	MathVista
Qwen2-VL-7B	64.8	60.7	54.1	58.2
Qwen2-VL-7B-VISC	66.3	60.4	54.3	58.5

Table 2: After fine-tuning, the model based on the Qwen2-VL architecture shows comparable or slightly improved performance across four vertical vision-language benchmarks.

method enhances the model’s ability to perceive visual information without diminishing its general vision-language capabilities.

#### RQ5: How is the quality of synthesized data?

Quality assurance is particularly critical for our fully automated data synthesis process. We conducted a rigorous quality assessment by engaging three Ph.D. students with expertise in computer vision and natural language processing. They evaluated a stratified random sample of 200 instances, with each instance independently assessed by all annotators. The overall accuracy is calculated based on three metrics: focus accuracy, sub-question correctness, and final answer correctness. A sample is counted as a positive instance when all three metrics are unanimously assessed as correct by at least two annotators, or regarded as a negative instance. With the synthesized data attaining 97.5% overall accuracy (Fleiss’  $\kappa$  (Fleiss, 1971) = 0.637), these results robustly validate the method’s reliability across measurement dimensions.

## 5 Conclusion

This study addresses the challenges from multi-image scenarios through two key innovations: (1) the *Focus-Centric Visual Chain*, a structured reasoning paradigm that breaks down complex tasks into targeted sub-questions with explicit visual focus, and (2) a bottom-up framework for automated synthesis of reasoning data in the form of *Focus-Centric Visual Chain*. Through this methodology, a large-scale multimodal dataset of 150K instances featuring multi-image compositions and cascaded reasoning chains, systematically generated via our automated synthesis framework. Models fine-tuned on VISC-150K achieve consistent improvements across multiple multi-image benchmarks. Our work not only breaks performance ceilings in existing tasks but also establishes an advanced framework for data-driven visual reasoning, providing actionable pathways to resolve persistent bottlenecks on multi-image tasks.



## Limitations

While our approach demonstrates promising results, we identify several important limitations. The *Focus-Centric Data Synthesis* framework requires pairwise relevance annotation across images, leading to quadratic computational complexity. We maintain moderate sizes of image sets to balance data diversity and computational efficiency.

Moreover, Our VISC-150K dataset primarily focuses on real-world photographs and comics. The approach’s effectiveness remains untested on structured visual content such as charts, diagrams, and code snippets, which may require different reasoning patterns.

In addition, the current implementation is constrained by the language models’ inherent capabilities. Consequently, our method inherits their limitations in managing complex spatial dynamics, domain-exclusive contexts, and expertise-dependent visual subtleties.

These limitations suggest promising directions for future research in multi-image understanding and reasoning.

## Acknowledgments

This work is also supported by the Public Computing Cloud, Renmin University of China and by fund for building worldclass universities (disciplines) of Renmin University of China. Yuhan Liu is supported by the "Qiushi Academic-Dongliang" Project of Renmin University of China (No. RUC24QSDL015).

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. [Openflamingo: An open-source framework for](#)

[training large autoregressive vision-language models](#). *arXiv preprint arXiv:2308.01390*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *arXiv preprint arXiv:2308.12966*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. [Graph of thoughts: Solving elaborate problems with large language models](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Webb. 2024. [Understanding the limits of vision language models through the lens of the binding problem](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 113436–113460. Curran Associates, Inc.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. [Are we on the right way for evaluating large vision-language models?](#) In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024c. [How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites](#). *Science China Information Sciences*, 67(12):220101.

Chuanqi Cheng, Jian Guan, Wei Wu, and Rui Yan. 2024. [From the least to the most: Building a plug-and-play visual reasoner via data synthesis](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4941–4957, Miami, Florida, USA. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [InstructBLIP: Towards general-purpose vision-language models with instruction tuning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

- Maryam Daniali and Edward Kim. 2023. [Perception over time: Temporal dynamics for robust image understanding](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5656–5665.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76(5):378–382.
- Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. [Blink: Multimodal large language models can see but not perceive](#). In *European Conference on Computer Vision*, pages 148–166. Springer.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. [G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model](#). *arXiv preprint arXiv:2312.11370*.
- Google Gemini Team. 2024. [Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context](#). *arXiv preprint arXiv:2403.05530*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoub, 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 (ICLR)*.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Bjorck, Vishrav Chaudhary, Subhojit Som, XIA SONG, and Furu Wei. 2023. [Language is not all you need: Aligning perception with language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 72096–72109. Curran Associates, Inc.
- Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. 2024. [Mantis: Interleaved multi-image instruction tuning](#). *Transactions on Machine Learning Research*.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. [What matters when building vision-language models?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 87874–87907. Curran Associates, Inc.
- Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. [Multimodal reasoning with multimodal knowledge graph](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10767–10782, Bangkok, Thailand. Association for Computational Linguistics.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. [LLaVA-onevision: Easy visual task transfer](#). *Transactions on Machine Learning Research*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. [Mvbench: A comprehensive multi-modal video understanding benchmark](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.
- Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. [Vila: On pre-training for visual language models](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26689–26699.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. [LLaVA-NeXT: Improved reasoning, OCR, and world knowledge](#).
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024c. [From skepticism to acceptance: simulating the attitude dynamics toward fake news](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*.
- Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2024d. [The step-wise deception: Simulating the evolution from true news to fake news with llm agents](#). *arXiv preprint arXiv:2410.19064*.

- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). In *The Twelfth International Conference on Learning Representations*.
- Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. [MMIU: Multimodal multi-image understanding for evaluating large vision-language models](#). In *The Thirteenth International Conference on Learning Representations*.
- OpenAI. 2024a. [GPT-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- OpenAI. 2024b. [OpenAI o1 System Card](#).
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. [Kosmos-2: Grounding multimodal large language models to the world](#). *Preprint*, arXiv:2306.14824.
- Yichong Shi, Yucheng Li, Yifan Zhang, Yizhuo Zhang, Yuchen Zhang, Yixin Wang, Yining Li, Zihan Zhang, Yizhuo Wang, Yixin Zhang, et al. 2024. [MathLLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- Alane Suhr, Stephanie Zhou, Iris Zhang, Huajun Zhang, Rui Bai, Yichen Xiong, and Yoav Artzi. 2019. [A Corpus for Reasoning about Natural Language Grounded in Photographs](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.
- Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024a. [Muirbench: A comprehensive benchmark for robust multi-image understanding](#). *arXiv preprint arXiv:2406.09411*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *arXiv preprint arXiv:2409.12191*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.
- Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024c. [Exploring the reasoning abilities of multimodal large language models \(mllms\): A comprehensive survey on emerging trends in multimodal reasoning](#). *Preprint*, arXiv:2401.06805.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. [Llava-cot: Let vision language models reason step-by-step](#). *arXiv preprint arXiv:2411.10440*.
- Huanjin Yao, Jiaying Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024. [Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search](#). *arXiv preprint arXiv:2412.18319*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. [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*, pages 9556–9567.
- Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. 2024a. [Native: Multi-modal knowledge graph completion in the wild](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, page 91–101, New York, NY, USA. Association for Computing Machinery.
- Yizhuo Zhang, Yifan Zhang, Yuchen Zhang, Yixin Wang, Yining Li, Zihan Zhang, Yizhuo Wang, Yixin Zhang, Yucheng Li, Yichong Shi, et al. 2024b. [MAVIS: Mathematical Visual Instruction Tuning](#). *arXiv preprint arXiv:2407.08739*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. 2024c. [Multi-modal chain-of-thought reasoning in language models](#). *Transactions on Machine Learning Research*.
- Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. [Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning](#). *arXiv preprint arXiv:2406.12742*.

## A Dataset Details

VISC-150K comprises 152,061 entries, with images aggregated from publicly accessible websites<sup>5</sup> established open-source visual datasets. It includes reasoning paths of lengths 1–8, formatted as open-ended QA and single-choice tasks, and covers varying numbers of input images. The detailed composition is illustrated in Figure 6.

## B Experimental Settings

All experiments were conducted on 8 NVIDIA A100 (80GB) GPUs. For training, we applied LoRA fine-tuning with a rank of 16, an initial learning rate of  $1e-5$ , a warmup ratio of 0.05, and a cosine learning rate scheduler. The batch size was set to 8 with a maximum context length of 32,768, using the bfloat16 floating-point format. During inference, we configured the temperature to 0 for deterministic generation, set max new tokens to 1,024, and accelerated computations with FlashAttention-2. Notably, for Qwen2-VL, we dynamically resized input image resolutions to the range of  $128 \times 28 \times 28$  –  $1280 \times 28 \times 28$  to balance inference speed and computational resource consumption.

## C Baseline Details

**Qwen2-VL** (Wang et al., 2024b) incorporates M-RoPE to handle positional information and is capable of processing images of any resolution dynamically. These techniques allow the model to capture detailed visual information more effectively.

**Qwen-VL-Chat** (Bai et al., 2023) is built upon the foundation of Qwen and is enhanced with visual capabilities through training on high-resolution images and fine-grained datasets.

**LLaVA-OneVision** (Li et al., 2025) is built upon LLaVA-NeXT series, leveraging large language models (LLMs) and vision encoders connected through a simple projection module.

**LLaVA-1.6** (Liu et al., 2024b) addresses the limitations of existing VLMs that primarily focus on single-image tasks. The model leverages an interleaved data format as a general template to unify different visual scenarios.

**LLaVA-1.5** (Liu et al., 2024a) is designed to improve visual reasoning and instruction-following capabilities by leveraging visual instruction tuning.

**InternVL2.5** (Chen et al., 2024b) is a state-of-the-art VLM built upon the architecture of InternVL2.

<sup>5</sup>WikiHow (<https://www.wikihow.com>), eHow (<https://www.ehow.com>)

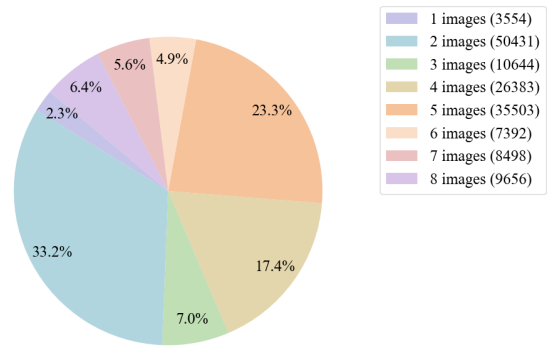


Figure 6: The image count distribution in VISC-150K spans 1–8 images per instance, with samples containing 2–5 images accounting for 80.9% of the total dataset.

Its key features include a robust vision encoder, flexible language model integration, dynamic high-resolution processing, and effective test-time scaling.

**InternVL2** (Chen et al., 2024c) employs a progressive alignment training strategy, allowing the model to scale from smaller to larger sizes while refining the training data from coarse to fine.

**Mantis-Idetics2** (Jiang et al., 2024) builds on existing LMM architectures and modifies them to support interleaved multi-image inputs. It uses a text-image interleaving format to mark boundaries between images, enabling the model to better understand and reason over multiple images.

**Idetics2** (Laurençon et al., 2024) incorporates learned pooling strategies to reduce the number of visual tokens, significantly improving computational efficiency while maintaining or even enhancing performance.

**VILA-1.5** (Lin et al., 2024) is designed to effectively integrate visual inputs with the strengths of large language models. Interleaved training data and joint SFT enable the model to achieve superior performance on vision-language tasks while retaining strong text-only capabilities.

**OpenFlamingo-v2** (Awadalla et al., 2023) utilizes frozen language models augmented with layers that cross-attend to outputs from a frozen vision encoder. The training on web-scraped image-text sequences enables it to process interleaved sequences of images and text.

## D Benchmark Details

**MMIU** (Meng et al., 2025) categorizes multi-image relationships into three primary types: se-

semantic, spatial, and temporal, grounded in cognitive psychology theory. These categories are further subdivided into seven subtypes, covering 52 distinct multi-image understanding tasks. The benchmark comprises 77K images and 11K multiple-choice questions.

**MuirBench** (Wang et al., 2024a) consists of 11,264 images and 2,600 multiple-choice questions, covering 12 distinct multi-image understanding tasks and encompassing 10 types of multi-image relationships. By incorporating diverse tasks and image relationships, it establishes a novel and comprehensive benchmark for multi-image understanding.

**MIRB** (Zhao et al., 2024) includes four evaluation dimensions: perceptual understanding, visual world knowledge integration, complex reasoning, and multi-hop reasoning. Each dimension incorporates specialized tasks requiring sophisticated cross-image comparison and analytical reasoning.

**BLINK** (Fu et al., 2024) reinterprets traditional computer vision problems as multiple-choice questions, incorporating 14 visual perception tasks that humans can quickly solve. It consists of 3,807 multiple-choice questions accompanied by 7,358 images sourced from multiple datasets, covering indoor, outdoor, and natural scenes.

**NLVR2** (Suhr et al., 2019) focuses on natural language grounding in visual contexts, presenting paired images with corresponding English descriptions. This benchmark emphasizes the evaluation of fine-grained linguistic-visual alignment through diverse reasoning tasks, requiring precise understanding of both textual and visual modalities.

**Mantis-Eval** (Jiang et al., 2024) comprises 217 high-quality multi-image reasoning samples, covering a range of multi-image skills such as coreference, reasoning, and comparison.

**MVBench** (Li et al., 2024) consists of 20 challenging video understanding tasks that assess temporal reasoning capabilities.

## E Prompt

For the four stages of data synthesis, we meticulously crafted structured prompts to achieve the finest-grained task decomposition at each phase. The prompts for each component are demonstrated below:

### Feature Extraction:

You are a visual description expert. Please provide a detailed, comprehensive, and natural language description of the following image, covering every visible detail.

**Overall View:** - Summarize the scene in 1-2 sentences, focusing on the general setting, lighting, time of day, and the environment. Ensure to include the general mood and ambiance.

**Main Objects:** - For each key object, describe these aspects in fluent natural language: - What is it (e.g., a person, a car, a building)? - Quantity, color, size, shape, material, texture, and any distinctive features. - Where is it located (foreground, center, background)? - State/Function: Is it active or stationary? What is its function in the scene?

**Secondary Objects and Background:** - Describe smaller or less prominent objects and elements in the background. How do they relate to the main objects? Mention any supporting objects that add depth to the scene (e.g., objects on a table, items in the background, etc.).

**Object Interactions:** - Highlight interactions or relationships between objects (e.g., people talking, animals interacting). Describe any dynamic actions or static arrangements.

**Text:** - If there are text in the image, extract all the text and analyze its meanings.

**Atmosphere&Theme:** - Convey the mood or theme of the scene, using descriptive adjectives (e.g., lively, serene, chaotic). If unsure, use "seems to" to indicate speculation about the tone.

**Detailed Natural Language Description:** - Integrate all of the above details into a flowing, cohesive narrative. Ensure to describe every element in fine detail, maintaining clarity and logical structure. Avoid redundancy or skipping any visible detail.

**Pair Connection:**

You are a professional visual content analyst skilled in analyzing image pairs that exhibit clear correlations.

User will provide a set of structured descriptions corresponding to images. Based on these descriptions, you are required to analyze the images through an object-oriented or event-oriented approach to identify which image pairs are most strongly correlated. Specifically, you should focus on determining whether there are common objects or associated events/themes between the images. By evaluating the co-occurrence of objects or the relationships between events or topics, return the correlated image pairs as a tuple.

**Relevance Annotation:**

You are a professional visual content analyst skilled in analyzing relationships between image pairs, including temporal, spatial, and semantic connections.

User will provide you with two images. Please generate relationship annotations between them based on the following requirements:

Task Requirements:

1. Temporal Relationship: Identify if there is a clear sequence of events between Image A and Image B. - First, analyze whether the scenes or events in the two images represent a clear chronological order. - If there is a clear temporal sequence, describe the progression or transition between the two images, noting the overall process.

2. Spatial Relationship: Analyze if there are any spatial connections or changes in scene or object positions between Image A and Image B. - Check if both images depict the same scene or objects in similar layouts. - If shared objects or settings are present, compare their positions, orientations, or size differences in both images.

3. Semantic Relationship: Evaluate if there is a thematic, emotional, or causal connection between Image A and Image B. - Determine if the themes, emotional tones, or meaning presented in both images are consistent or related. - Assess if there is a cause-and-effect relationship or logical connection between the two images. Output Format should be in JSON.

Distribution	Samples
Unanimously Correct	191
2 Correct vs. 1 Incorrect	4
1 Correct vs. 2 Incorrect	2
Unanimously Incorrect	3

Table 3: Statistical analysis of human annotation results conducted on 200 sampled instances. Each instance’s annotation outcome was classified into four categories: (1) unanimously correct, (2) two annotators rated as correct vs. one incorrect, (3) two annotators rated as incorrect vs. one correct, and (4) unanimously incorrect. A data instance is considered valid if at least two annotators labeled it as correct; otherwise, it is deemed invalid.

**Quesiton Generation:**

Task Requirements: 1. Generate Three Complex Reasoning Questions: - Each question should be a multi-step reasoning question, and involve at least three images. - Questions should be object-oriented, or event-oriented. - Avoid begin with 'How' if possible, and make sure the answer is not open-ended. - Questions should be about fine-grained features instead of coarse understanding. - Questions types: - Detail analysis and comparison - Fact judgment - Sequence ordering - Scene understanding - Visual grounding - Counterfactual reasoning - Action prediction - Visual navigation - Don’t specify images explicitly. - Each question must be a single sentence without clauses connected by 'and'.

2. Decompose Each Complex Question into Sub-Questions and Build a Reasoning Chain: - Each sub-question specifies one or two images. - Don’t focus on the same image twice. - Construct a logical reasoning chain for each question, showing the step-by-step connection of sub-questions and answers.

3. Step-by-Step Answer Each Reasoning Chain to Arrive at the Final Answer

4. Ensure Data Quality: - The questions and answers must be clear, specific, and logically consistent. - Avoid irrelevant details or ambiguity, ensuring that all generated content is directly related to the provided image information. Output Format should be JSON.

**F Details of Human Annotations**

We recruited three annotators (Ph.D. students in Computer Science and Technology) to evaluate the

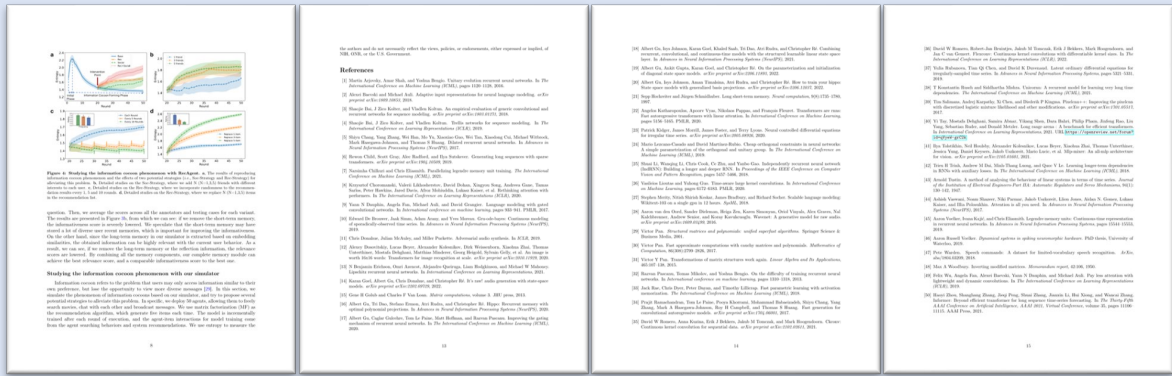
correctness of 200 randomly sampled instances from VISC-150K. For each instance, the evaluation criteria included three dimensions:

- Final Answer Correctness: Whether the annotator deemed the final answer to accurately resolve the original question;
- Sub-Question Answer Correctness: Whether intermediate answers correctly addressed their corresponding sub-questions;
- Visual Focus Validity: Whether the visual focus at each step aligned with the minimal essential visual information required for reasoning.

An instance was labeled as correct by an annotator only when all three criteria were satisfied. The annotation results from the three annotators are summarized in Table 3. A data instance was marked as valid if at least two annotators labeled it as correct; otherwise, it was deemed invalid. Based on the evaluation results, our synthetic data achieved a 97.5% validity rate under substantial inter-annotator agreement (Fleiss'  $\kappa = 0.637$ ).

## G Case Study

We present test cases for Qwen2-VL-VISC and LLaVA-OneVision-VISC fine-tuned on VISC-150K, illustrated in Figure 7 and Figure 8 respectively. These cases effectively demonstrate the multi-image reasoning and integrated information processing capabilities of our data-augmented models.



Question: What's the paper title of citation [21] on page 14?

Qwen2-VL-VISC:

Step\_1: Identify the citation [21] in page number. Focus: Image2,Image3,Image4\n Answer: Citation [21] is located on page 14.

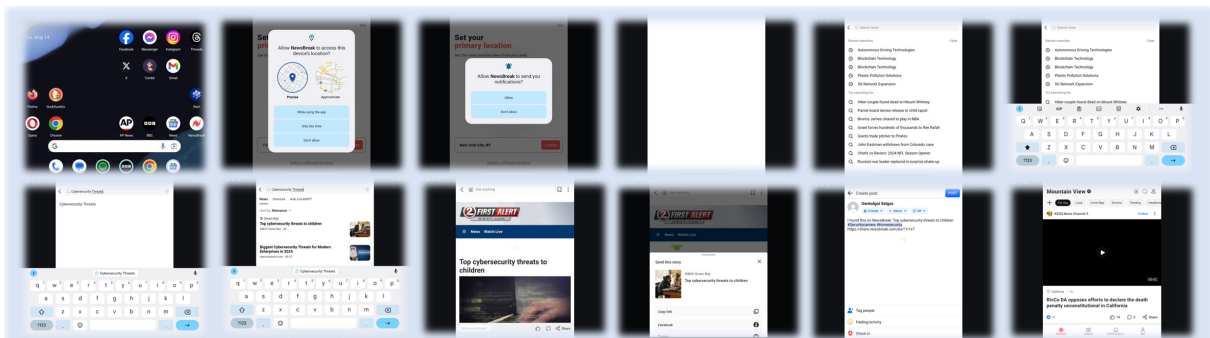
Step\_2: Determine the paper title referenced by citation [21].\n Focus: Image2\n Answer: The paper title referenced by citation [21] is "Long short-term memory".

Step\_3: Confirm the details of citation [21]. Focus: Image2\n Answer: Citation [21] is a study named "Long short-term memory", conducted by two authors and published on Neural computation.

Final\_answer: The paper title referenced by citation [21] is "Long short-term memory".

Figure 7: In the cross-image visual reasoning case, Qwen2-VL-VISC trained on the VISC-150K dataset with the acquired Focus-Centric Visual Chain reasoning paradigm accurately resolves the reasoning task.





The corresponding actions are: step 1: CLICK: (929, 676)\nstep 2: CLICK: (561, 612)\nstep 3: CLICK: (516, 514)\nstep 4: CLICK: (716, 74)\nstep 5: CLICK: (356, 85)\nstep 6: TYPE: Cybersecurity Threats\nstep 7: CLICK: (856, 874)\nstep 8: CLICK: (452, 314)\nstep 9: CLICK: (724, 915)\nstep 10: CLICK: (508, 873)\nstep 11: CLICK: (733, 79)\nstep 12: COMPLETE\n

Which app-combination list was used in this GUI navigation episode?

LLaVA-OneVision-VISC:

Step\_1:\n What is the first action taken in the screenshot?\n References: image1\n Answer: The first action taken is clicking on a button labeled 'AP News'.\n

Step\_2:\n What is the second action taken after clicking on the 'AP News' button?\n References: image2\n Answer: The second action taken is selecting 'NewsBreak' from the list of options.\n

Step\_3:\n What is the third action taken after selecting 'NewsBreak'?\n References: image3\n Answer: The third action taken is clicking on a location option to set the primary location.\n

Step\_4:\n What is the fourth action taken after setting the primary location?\n References: image4\n Answer: The fourth action taken is typing 'Cybersecurity Threats' into the search bar.\n

Step\_5:\n What is the fifth action taken after typing 'Cybersecurity Threats'?\n References: image5\n Answer: The fifth action taken is clicking on a news article titled 'Top cybersecurity threats to children'.\n

Step\_6:\n What is the sixth action taken after clicking on the news article?\n References: image6\n Answer: The sixth action taken is sharing the news article via social media platforms.\n

Step\_7:\n What is the seventh action taken after sharing the news article?\n References: image7\n Answer: The seventh action taken is creating a post on a social media platform with the shared news article link.\n

Step\_8:\n What is the eighth action taken after creating the post?\n References: image8\n Answer: The eighth action taken is viewing the shared news article on a different device or browser.\n

Step\_9:\n What is the ninth action taken after viewing the shared news article?\n References: image9\n Answer: The ninth action taken is scrolling through a news feed on a social media platform.\n

Step\_10:\n What is the tenth action taken after scrolling through the news feed?\n References: image10\n Answer: The tenth action taken is watching a video on a news channel.\n

Final\_answer: ['NewsBreak', 'Facebook']

Figure 8: In a cross-image GUI navigation case study, LLaVA-OneVision-VISC, guided by the *Focus-Centric Visual Chain* paradigm, constructs a multi-step reasoning trajectory. By iteratively focusing on GUI elements within individual images at each stage, it deduces a comprehensive operational workflow and ultimately arrives at the correct solution.