MMEvol: Empowering Multimodal Large Language Models with Evol-Instruct

Anonymous ACL submission

Abstract

The development of Multimodal Large Language Models (MLLMs) has seen significant progress, driven by increasing demands across various fields (e.g., multimodal agents, embodied intelligence). While model-driven approaches aim to enhance MLLM capabilities through diverse architectures, their performance gains have become increasingly marginal. In contrast, data-driven methods, which scale up image-text instruction datasets, have proven more effective but face challenges related to limited data diversity and complexity. The absence of high-quality instruction data remains a major bottleneck in MLLM development. To address this issue, we propose **MMEvol**, a novel multimodal instruction data evolution framework. This framework iteratively enhances data quality through a refined combination of fine-grained perception, cognitive reasoning, and interaction evolution, generating a more complex and diverse image-text instruction dataset that significantly improves MLLM capabilities. Starting with an initial dataset, SEED-163K, we employ **MMEvol** to systematically expand instruction diversity, extend visual reasoning steps to improve cognitive abilities, and extract fine-grained visual details to enhance understanding and robustness. To rigorously evaluate our approach, we conduct extensive qualitative analysis and quantitative experiments across 13 vision-language tasks. Compared to baseline models trained on the original seed dataset, our method achieves an average accuracy improvement of 3.4 percentage points. Moreover, our approach attains state-of-the-art (SOTA) performance in nine tasks while using significantly less data than existing state-of-the-art models.

1 Introduction

Multimodal Large Language Models (MLLMs) (Liu et al., 2024b; Li et al., 2023c; Dai et al., 2024; Luo et al., 2024) have advanced rapidly over the past two years, emerging as the dominant paradigm for vision-language tasks (Kembhavi et al., 2016; Fu et al., 2024; Zhang et al., 2024; Qian et al., 2024). These models integrate visual encoders (Radford et al., 2021; Zhai et al., 2023; Sun et al., 2023) with large language models (LLMs) (Touvron et al., 2023; Bai et al., 2023; Lu et al., 2024; Young et al., 2024; Tao et al., 2024), leveraging large-scale coarse-grained image-text pre-training (Zhu et al., 2024; Schuhmann et al., 2022, 2021), followed by small-scale instruction-tuning (Chen et al., 2024a; Liu et al., 2024b). This approach has led to remarkable advancements in vision-language understanding, enabling applications across multimodal agents, embodied intelligence, and real-world reasoning tasks. Existing model-driven approaches (Liu et al., 2024a; Tong et al., 2024) seek to enhance MLLM performance by refining network architectures for more effective integration of textual and visual knowledge. However, the marginal benefits of

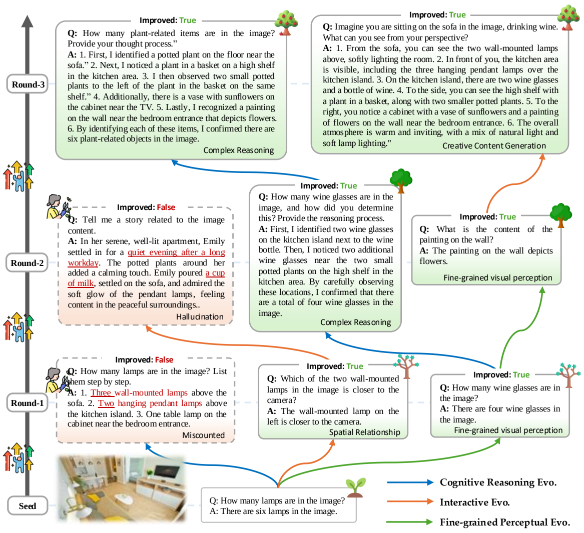


Figure 1: **Overview of MMEvol.** Instruction evolution and instruction elimination synergistically collaborate through multiple rounds to enhance the diversity and complexity of instruction data.

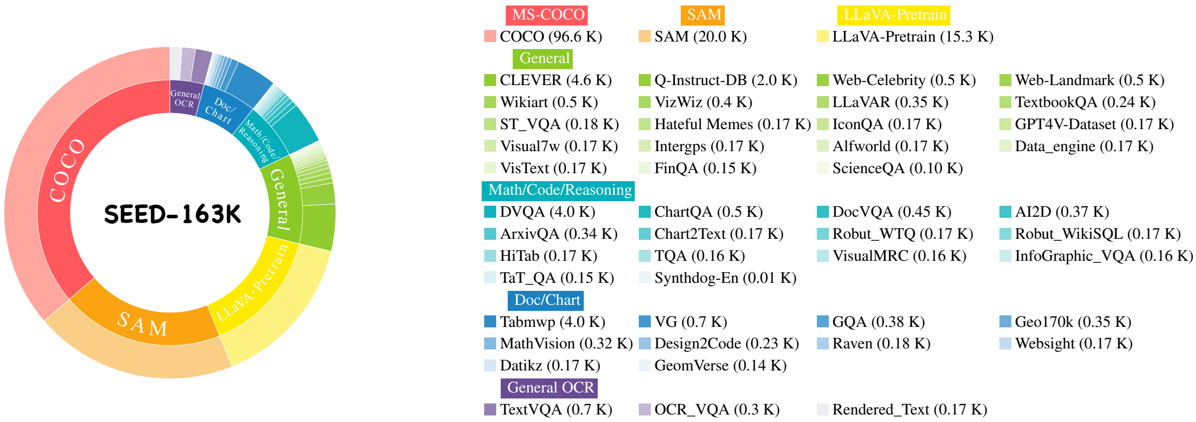


Figure 2: **SEED-163K: 163K Curated Seed Instruction Tuning Dataset for Evol-Instruct. Left:** The inner circle shows the original distribution of SEED-163K. The outer circle shows the curated SEED-163K. **Right:** All the data sources in the SEED-163K dataset, as well as the ones filtered in data curation.

these architectural improvements are diminishing due to the absence of high-quality training data. Furthermore, redundant architectural modifications fail to push the boundaries of model intelligence. Conversely, data-driven approaches (Liu et al., 2024b; Chen et al., 2024a; Yu et al., 2023; Liu et al., 2024c; Fang et al., 2024; Chen et al., 2023) have demonstrated greater effectiveness by expanding the scale of image-text instruction datasets. However, these methods often suffer from limited data diversity and insufficient complexity, constraining the models' ability to generalize and reason effectively. The lack of high-quality, diverse instruction data remains one of the most critical challenges in advancing MLLMs. Thus, there is an urgent need to develop automated methods that can generate more complex, diverse, and high-quality instruction data at a low cost, thereby enhancing MLLM capabilities.

An analysis of existing data-driven methods for generating image-text instruction data reveals three major limitations. First, limited instruction diversity. Manually annotated instructions are constrained by human cognitive limitations, while model-generated instructions rely on template-based presets. This restricts the diversity of task formulations, making it difficult for MLLMs to generalize to real-world scenarios and limiting their instruction-following capabilities. Second, limited instruction complexity. Manually annotated instructions tend to be simple or moderately complex, while automatically generated instructions are often brief and lack multi-step visual reasoning. This prevents MLLMs from effectively handling complex vision-language tasks that require deeper cognitive reasoning. Third, insufficient alignment granularity. Both manually and model-generated instructions primarily focus on common objects, often neglecting rare or small visual elements. This results in coarse-grained image-text alignment, reducing the model's visual perception robustness and increasing susceptibility to hallucinations.

To overcome these limitations, we propose **MMEvol**, a novel multimodal instruction evolution framework that leverages advanced MLLMs for iterative instruction refinement. This framework automatically generates diverse, open-domain instructions at multiple difficulty levels, enhancing the overall instruction-following capabilities of MLLMs. Given that visual-language instruction data must be tightly aligned with image content, existing Evol-Instruct (Xu et al., 2023; Luo et al., 2023a,b) methods often produce redundant restatements or instructions unrelated to images, making deep and broad instruction evolution difficult. To address this, we introduce a refined image-text instruction paradigm, which defines three key evolution directions. i) Fine-grained perception evolution enhances visual detail extraction, generating instructions that capture more granular information from images. ii) Cognitive reasoning evolution extends multi-step reasoning chains, improving the model's ability to process complex visual tasks. iii) Interaction evolution increases instruction diversity by generating a wider variety of instruction formats.

The **MMEvol** mechanism is summarized in fig. 1, where each evolution cycle consists of two main steps: instruction evolution and instruction elimination. During instruction evolution, one of the three evolution directions is randomly selected to transform simple instructions into more complex or diverse forms. Since not all instruction transformations are successful, an instruction elimination step is applied to filter out ineffective or redundant evolutions. By iteratively refining the dataset through multiple rounds of evolution and elimination, **MMEvol** produces a high-quality, complex instruction dataset that enhances MLLM performance across a broad range of tasks.

To validate the effectiveness of **MMEvol**, we perform three rounds of evolutionary iterations on 163K seed data, leading to 447K evolved samples. We fine-tuned the open-source LLaVA-NeXT (Liu et al., 2024a) model with these evolved data and compared it with other advanced methods across 13 vision-language benchmarks. Our method achieves state-of-the-art (SOTA) performance, demonstrating the effectiveness and efficiency of **MMEvol**. Additionally, we conduct detailed qualitative analysis and ablation experiments to showcase the contribution of each component of our method. We hope that the released evolutionary data and code will assist the community in understanding that using a small amount of high-quality image-text instruction data is far more critical than training MLLMs with large-scale low-quality image-text instruction data.

Our main contributions can be summarized as follows: (1) A image-text instruction evolution framework, **MMEvol**, is designed to leverage advanced MLLMs, automating the generation of open-domain image-text instruction data across varying difficulty levels to enhance the diversity and complexity of existing datasets. (2) By utilizing instruction evolution data, a high-quality data recipe is composed, and the evolved data will be released to advance the capabilities of other open-source MLLMs further. (3) We train an MLLM using this high-quality data recipe, achieving superior performance in various downstream visual-language tasks compared to other fully open-source methods.(4) The effectiveness and efficiency of the proposed approach are validated through extensive qualitative and quantitative analyses.

2 Related Work

Multimodal Large Language Models (MLLMs). MLLMs have experienced rapid advancements in recent years, driven by the success of Large Language Models (LLMs) and the increasing availability of diverse image-text instruction data sourced from the internet. LLaVA (Liu et al., 2024b) and MiniGPT-4 (Zhu et al., 2023) have demonstrated strong cross-task generalization by integrating visual encoders with large language models through lightweight connectors and training on instruction datasets. LLaVA-NeXT (Liu et al., 2024a) has significantly enhanced visual perception by employing dynamic resolution techniques, while Cambrain-1 (Tong et al., 2024) has improved model robustness via visual encoder routing, though at the cost of higher training overhead. In contrast, DEEM (Luo et al., 2024) simplifies MLLM architectures and improves model robustness by leveraging diffusion models to extract visual features, replacing traditional visual encoders. Subsequent work (Wang et al., 2024b; Zhou et al., 2024; Xie et al., 2024) builds upon DEEM, combining diffusion models with LLMs to further enhance generative and comprehension capabilities. However, despite these advances, MLLMs still face major challenges related to data quantity and quality, which constrain further performance improvements.

Image-text Instruction Data Construction. LLaVA (Liu et al., 2024b) has advanced MLLM capabilities by introducing LLaVA-Instruct (Liu et al., 2024b), a dataset labeled by advanced LLMs. However, this approach fails to fully exploit visual information and offers a limited variety of instruction types. ALLaVA (Chen et al., 2024a), which relies on manual crafting and rewriting of instruction data, introduces greater diversity but suffers from high annotation costs, inefficiency, and overly simplistic task formulations. To address these limitations, MMInstruct (Liu et al., 2024c) generates instruction data automatically using advanced MLLMs, but its instruction complexity and diversity remain constrained by predefined formats, preventing the full utilization of visual content. Similarly, VILA^2 (Fang et al., 2024) has explored instruction evolution to generate extensive datasets, yet it still lacks sufficient complexity and variation, limiting its broader applicability. In contrast, we introduce **MMEvol**, a novel instruction evolution framework that iteratively enhances both instruction diversity and complexity using limited data. By extracting richer visual information and refining instructional structures, **MMEvol** equips MLLMs with more powerful multimodal reasoning and comprehension capabilities.

3 Method

In this section, we first describe the curation process for the seed instruction data and then detail the methodological framework of **MMEvol**.

Due to space constraints, we provide a concise overview of the seed data curation process and prompt templates, with additional details available in appendix E.

3.1 Seed Data Curation

The seed instruction data are curated from LLaVA-Instruct (Liu et al., 2024b) and ShareGPT4V (Chen et al., 2023), supplemented with scientific and chart-based data sampled from Cambrain-1 (Tong et al., 2024). This process involves careful selection and refinement to ensure high data quality and diversity. For caption-only instructions, we employ the OpenAI GPT-4o mini API to generate structured instruction data. Following merging and filtering, we obtain a comprehensive dataset containing 163K instruction samples, each paired with a unique image. This dataset serves as the foundation for our Evol-Instruct framework. The composition of the seed dataset is illustrated in fig. 2. Additional details on the seed data curation process are provided in appendix A.

3.2 Methodological Details

The evolution of image-text instruction data is inherently constrained by visual information, requiring evolved instruction data to maintain strong alignment with image content to prevent hallucinations. This makes diversity evolution in image-text instructions particularly challenging, as models struggle to introduce meaningful variations without deviating from image-grounded content. Additionally, increasing the complexity of instruction data often leads to shallow reasoning phenomena, where MLLMs fail to generate deep, structured responses.

As shown in fig. 1, we address these challenges by designing a structured evolution framework that incorporates four key domains: visual objects, atomic capabilities, visual manipulations, and instruction formats. These domains standardize instruction data to ensure both quality and diversity during evolution. The visual object domain explicitly includes objects present in the image, constraining instruction evolution to image-relevant content, thereby reducing hallucinations. We further define an atomic capability domain, which consists of nine key capabilities for image-text reasoning. This includes five vision-centric capabilities—localization, reference, computation, optical character recognition (OCR), and existence judgment—and four language-centric capabilities—relation description, scene understanding, behavior prediction, and world knowledge association. These atomic capabilities enhance data diversity, enabling models to process a broader range of multimodal tasks. To mitigate shallow reasoning, we introduce a visual manipulation domain, which structures problem-solving as multi-step manipulation chains. Each step explicitly follows a vision-centric atomic capability, ensuring that the visual reasoning process remains structured and interpretable. Finally, the instruction format domain defines various interaction types, enabling models to handle a diverse range of task formulations. Together, these adaptations enhance the diversity and complexity of image-text instruction data, improving the overall effectiveness of instruction evolution.

Fine-grained Perceptual Evolution. The goal of fine-grained perceptual evolution is to maximize the extraction of detailed visual information from images, with a specific focus on overlooked, non-primary visual objects. In existing datasets, most instructions center around prominent objects, while less frequent, long-tail objects are often ignored. This lack of diversity results in suboptimal model generalization, leading to hallucinations and reduced robustness when encountering uncommon visual elements. To address this issue, fine-grained perceptual evolution generates instructions that introduce new visual objects, expanding the dataset's coverage of rare and underrepresented elements. By uncovering previously overlooked but visually meaningful information, this approach enhances both visual grounding and model robustness. The evolutionary prompt template and process are illustrated in fig. 3.

Cognitive Reasoning Evolution. Reasoning ability is a fundamental capability of multimodal large language models (MLLMs). However, existing instruction datasets, such as LLaVA-Instruct (Liu et al., 2024b), primarily consist of simple question-answer pairs that lack detailed reasoning steps. This limits the ability of trained models to handle complex reasoning tasks, such as multimodal agent interactions and visual reasoning. To address this, we introduce the concept of a visual manipulation chain, which abstracts four vision-centric reasoning capabilities into structured visual operations. These operation functions, described in text, define a structured process for incremental reasoning in multimodal tasks. By explicitly generating and incorporating multi-step visual reasoning, we establish a scalable framework for defining instruction

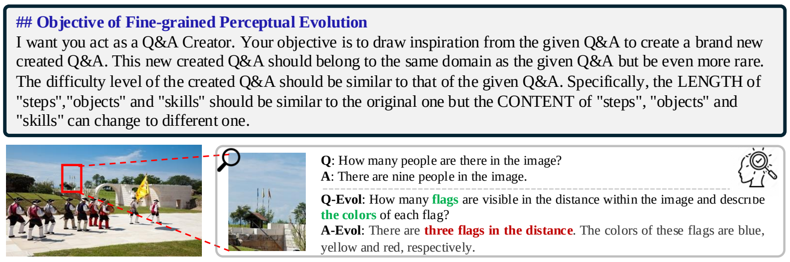


Figure 3: **Fine-grained perceptual evolution prompt and example.** Fine-grained perceptual evolution can generate samples with more detailed visual information, enhancing data diversity, which are marked with different colors for better visualization.

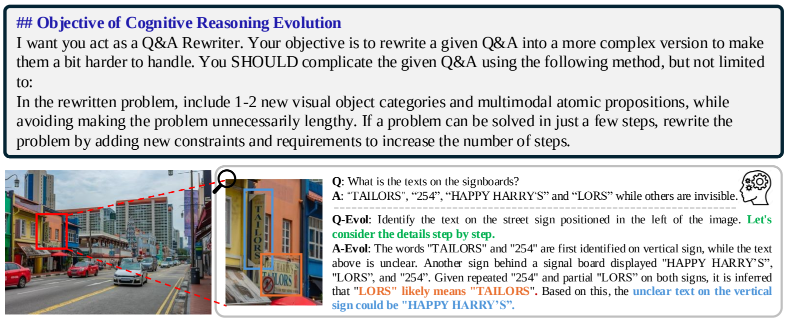


Figure 4: **Cognitive reasoning evolution prompt template and example.** Cognitive reasoning evolution can endow instruction data with a longer visual reasoning chain, increasing the complexity of the data. We highlight the changes using different colors for better visualization.

complexity. During the cognitive reasoning evolution process, we iteratively enhance instruction data by increasing the depth of visual reasoning steps, resulting in more complex and nuanced instructions. This evolution allows MLLMs to develop a richer understanding of visual concepts and reasoning patterns. The evolutionary prompt template and process are illustrated in fig. 4.

Interactive Evolution. Most existing models generate instruction data in only a few predefined formats. For example, LLaVA-Instruct primarily supports dialogue-based question-answering, complex reasoning, and global description tasks. Similarly, handcrafted datasets, such as ALLaVA (Chen et al., 2024a), are constrained by annotators' expertise, limiting the range of instruction formats and making it challenging to design diverse task structures. Models trained on such limited instruction formats often struggle to follow complex and varied user-specified instructions, reducing their practical applicability in real-world multimodal interactions. To overcome this limitation, we propose interactive evolution, an approach that automatically generates diverse instruction forms, enriching the range of interaction experiences for MLLMs. This method ensures that models are trained on a wider spectrum of instruction formats, improving their ability to handle real-world multimodal queries. The evolutionary prompt template and process are illustrated in fig. 5.

Instruction Elimination. After each evolutionary cycle, we evaluate the evolved instruction data across multiple dimensions to assess the effectiveness of the evolution process. Instruction data that demonstrate evolutionary gains are retained, while those that fail to meet the expected improvements are discarded. This selective process ensures that only high-quality, refined instruction data contribute to the training of MLLMs. The evolutionary elimination prompt template and process are illustrated in fig. 6.

4 Experiments

4.1 Benchmarks

To comprehensively evaluate the effectiveness of our evolutionary method, we select 13 benchmarks, with their sources and tested skills detailed in table 11. In addition to the classic visual language tasks, we introduce several more cutting-edge tasks to more comprehensively assess the quality of our data. MIA (Qian et al., 2024) is an open-domain instruction-following benchmark designed to rigorously assess a model's instruction adher-

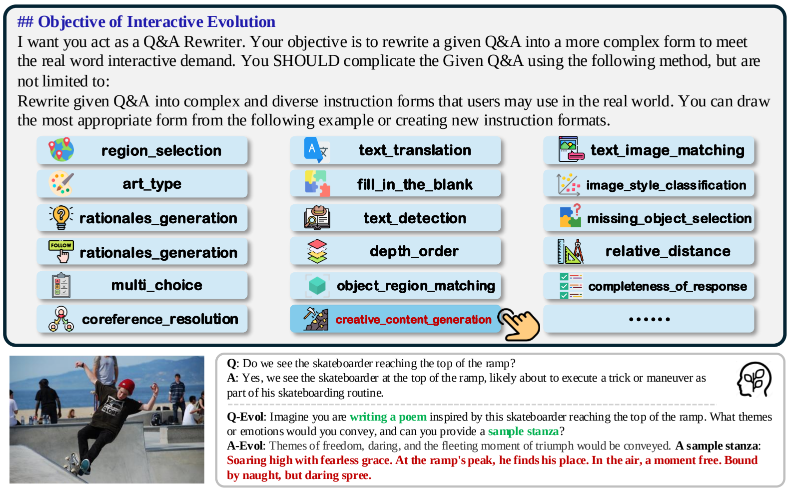


Figure 5: **Interactive evolution prompt template and example.** Interactive evolution can automatically generate various types of non-predefined instruction formats, significantly enhancing the diversity of the data. The differences are highlighted using distinct colors for better visualization.

Instruction Elimination
Evaluate the difficulty and complexity of each rewritten samples compared to its original ones. Determine if the rewritten sample has improved in quality via "yes/no”.
Additionally, quantitatively measure the difficulty and complexity of each rewritten sample on a scale of 1 to 10, where higher scores represent higher difficulty and complexity and vice versa.
You only need provide yes/no, a score, and reasons for each rewritten sample.

{“improved”: “yes/no”, “score”: 1-10, “reason”: the reason for the improvement and score},

Evaluation Criteria
- Length: Longer Q&A pairs generally have more detail and thus are considered more complex.
- Semantic Complexity: Use of more sophisticated language or concepts.
- Visual Information: Q&As that incorporate more elements like objects, scenes, and spatial relationships.
- Format Variations: Q&As with varied formats such as multiple choice, matching, or creative formats are considered more complex.
- Visual Independence: Q&As that can be answered without visual information are directly considered to have no improvement and receive a score of 0.
Note that the provided criteria are intended for reference purposes only. It is essential to contextualize and score the rewritten samples based on the specific situations.

In-context QA samples with different difficulties (1-10)

Figure 6: **Instruction elimination prompt template.** Instruction elimination is used to calculate the evolutionary gain and complexity level of the instruction data. We filter out harmful data that failed to evolve based on the evolutionary gain.

ence using a diverse set of instructions. MM-Self-Instruct (Zhang et al., 2024) is a visual reasoning benchmark that evaluates a model's visual perception capabilities, focusing on common visual reasoning tasks encountered in real-world scenarios.

4.2 Implementation Details

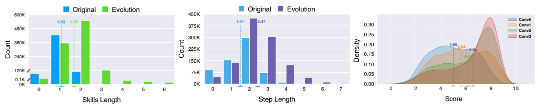
Data. During pre-training, we use LLaVA-Pretrain-595K (Liu et al., 2024b) for image-text alignment training. To ensure a fair comparison, we conduct ablation experiments by fine-tuning models separately on seed data and evolved data, allowing us to directly assess the benefits of **MMEvol**. For state-of-the-art (SOTA) comparisons, we fine-tune models using evolved instruction data in combination with other publicly available datasets, including samples from Cambrain-1 (Tong et al., 2024), and compare the results against existing methods. Further details on training data configurations can be found in appendix B.

Model. We adopt the LLaVA-NeXT architecture, where a multimodal large model consists of three primary components: i) A large language model (LLM) for next-token prediction; ii) A visual encoder for extracting visual features; iii) An image-text projector for aligning visual and textual modalities. For ablation experiments, we use Llama3-8B-Instruct (Touvron et al., 2023). For comparisons with other methods, we utilize our previous SOTA settings, employing Llama3-8B-Instruct and Qwen2-7B-Instruct (Bai et al., 2023). We adopt CLIP-ViT-L (Radford et al., 2021) as the visual encoder, with simple linear layers serving as the bridge between image and text modalities.

Training Strategies. We train **MMEvol** following the widely used two-stage paradigm: i) Vision-Language Pre-training; ii) Visual Instruction-Tuning. The language model and ViT are pre-trained separately, while the projector is randomly initialized. To initially align the visual and text feature spaces, we utilize a pre-aligned dataset. Finally, we perform instruction tuning using vision-language instruction datasets. All experiments are conducted on 8 × A100 GPUs, with a global batch size of 128. We employ the AdamW optimizer (Loshchilov, 2017), setting learning rates to 5×10^{-5} and 2×10^{-5} for the two training stages, respectively. Each stage is trained for one epoch using a 3% warm-up strategy. Additional training details are provided in appendix B.

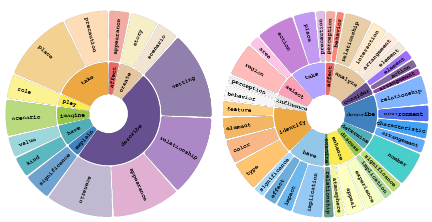
FP-Evol	I-Evol	CR-Evol	I-Elim	MMStar	MathVista^{M}	POPE	AI2D	MME^{C}	MMMU^{V}	RWQA	AVG.
✗	✗	✗	✗	36.5	25.3	84.8	53.9	31.5	32.3	43.5	44.0
✔	✗	✗	✗	37.3 (+0.8)	25.6 (+0.3)	85.0 (+0.2)	54.2 (+0.3)	33.0 (+1.5)	32.5 (+0.2)	46.7 (+3.2)	44.9 (+0.9)
✔	✔	✗	✗	38.2 (+1.7)	26.2 (+0.9)	83.8 (-1.0)	54.5 (+0.6)	35.6 (+4.1)	32.9 (+0.6)	48.9 (+5.4)	45.8 (+1.8)
✔	✔	✔	✗	38.9 (+3.4)	27.3 (+3.0)	83.6 (-1.2)	54.7 (+0.8)	**40.1** (+8.6)	34.4 (+0.9)	54.4 (+10.9)	47.6 (+3.8)
✔	✔	✔	✔	**40.3** (+3.8)	**28.6** (+3.6)	**86.5** (+1.7)	**55.2** (+1.3)	39.9 (+8.4)	**35.3** (+3.0)	**55.3** (+11.8)	**48.7** (+4.7)

Table 1: **Ablation study on instruction evolution and instruction elimination (6K).** Ablation study on instruct. Instruction evolution enhances data complexity and diversity, while the integration of instruction elimination further refines data quality, significantly reducing the occurrence of visual hallucinations.



(a) (b) (c)

Figure 7: (a) The skills length distribution between the seed data and our evolved data; (b) The reasoning steps length distribution between the seed data and our evolved data; (c) The difficulty and complexity level distribution between the seed data and our evolved data.



(a) Seed data (b) Evolved data

Figure 8: The root verbs (inner circle) and their top noun objects (outer circle) of the seed data in (a) and the evolved data in (b).

4.3 Qualitative Analysis

We randomly sample 30K data points from the seed dataset and perform a qualitative analysis of the instruction data before and after evolution. As illustrated in fig. 7, the evolved instruction data exhibits significantly greater complexity. Specifically, each evolved instruction incorporates 0.68 more atomic abilities, as shown in fig. 7a, and has an average visual operation chain reasoning length that is 0.86 steps longer compared to pre-evolution data in fig. 7b. Furthermore, as depicted in fig. 7c, the average difficulty score increases progressively across evolutionary rounds, demonstrating the effectiveness of cognitive reasoning evolution in enhancing instruction complexity.

To analyze the types and diversity of evolved instructions, we extract verb-noun structures from the generated instruction data. We utilize the Berkeley Neural Parser (Kitaev and Klein, 2018; Kitaev et al., 2018) to parse each instruction, extracting the root verb and its first direct noun object. The distribution of root verbs and their noun objects, with occurrences exceeding 2K, is visualized in fig. 8. The results indicate that evolution significantly enhances instruction diversity, leading to richer intents and varied textual formats. More qualitative analysis can be found in appendix C.

4.4 Ablation Study

We conduct ablation studies on seven vision-language benchmarks to analyze the impact of instruction evolution and elimination. As presented in table 1, the different evolution processes can be orthogonally superimposed, progressively enhancing data diversity and complexity, resulting in an average performance gain of 3.8 points across multiple benchmarks. However, when instruction elimination is omitted, failed evolutions introduce harmful data, leading to a 1.2-point reduction in the model's resistance to hallucinations, as measured on POPE (Li et al., 2023d). When both instruction evolution and elimination are applied, instruction elimination effectively filters out harmful data from failed evolutions, improving the quality and density of the evolved dataset. This results in an additional performance gain of 0.9 points on average, with hallucination resistance improving by 1.7 points, aligning with the qualitative analysis results presented in section 4.3. Additional ablation studies are detailed in appendix D.

4.5 Benchmark Comparison

After comprehensively validating our approach's ability to enhance instruction data complexity and diversity, we conduct a thorough comparison with previous SOTA methods across 13 vision-language benchmarks, with results summarized in table 2. Supported by enhanced and refined instruction data, our MLLM significantly outperforms existing models across almost all benchmarks, aligning with the performance improvements observed in our ablation experiments in section 4.4. Notably, compared

Model	PT	IT	VQA^{v2}	GQA	MME^{C}	MMStar	HallBench	MathVista^{M}	MMMU^{V}	AI2D	POPE	MIA	BLINK	RWQA	MMSInst	AVG.
Weight Open-Source																
Yi-VL-6B	125M	1M	–	–	46.2	37.7	55.7	28.8	40.3	59.8	82.5	26.1	38.7	53.5	–	46.9
DeepSeek-VL-7B	275M	50M	–	–	37.1	40.5	53.9	36.8	38.3	65.3	85.6	61.0	40.9	49.7	26.7	48.7
Qwen-VL-Chat-7B	1.4B	50M	78.2	57.5	49.0	34.5	56.4	34.9	37.0	63.0	74.9	63.1	28.2	49.3	–	52.2
CogVLM-Chat-17B	1.5B	5.1M	–	65.2	37.4	39.9	55.1	34.7	37.3	63.3	88.0	60.0	41.5	60.3	–	53.0
MiniCPM-V2.5-8B	570M	9.1M	81.9	64.7	50.3	51.3	59.2	**54.3**	43.0	**78.3**	86.7	76.3	36.7	63.5	28.2	59.6
Fully Open-Source																
InstructBLIP-7B	0.6M	0.8M	–	49.2	31.8	32.7	53.6	24.4	30.6	40.6	86.1	38.2	39.7	36.9	–	42.2
LLaVA-1.5-7B	0.6M	0.7M	78.5	62.0	37.8	33.1	48.8	25.6	35.7	55.5	86.1	62.2	38.0	54.8	15.4	48.7
LLaVA-1.5-13B	0.6M	0.7M	80.0	63.3	34.8	34.3	45.3	27.7	37.0	61.1	88.4	63.6	40.9	55.3	–	52.6
LLaVA-NeXT-8B	0.6M	0.8M	81.8	65.2	44.6	43.9	52.3	31.5	41.7	69.9	87.3	65.1	43.5	60.1	25.6	54.8
LLaVA-NeXT-13B	0.6M	0.8M	82.8	65.4	37.1	40.4	51.5	35.1	35.9	72.2	87.8	69.2	41.2	59.1	30.2	54.5
VILA-1.5-8B	50.5M	6.0M	80.9	61.9	39.0	39.7	55.8	37.3	36.9	58.8	85.5	66.1	37.0	43.3	21.6	51.1
VILA-1.5-13B	50.5M	6.0M	82.8	64.3	38.5	44.2	59.2	42.5	37.9	69.9	84.2	61.2	41.5	53.3	30.6	54.6
Cambrian-1-8B	2.5M	7.0M	81.2	64.6	41.1	50.7	47.8	47.0	41.8	74.6	86.4	68.7	44.9	**64.2**	28.3	57.1
Cambrian-1-13B	2.5M	7.0M	82.6	64.3	44.5	47.1	58.9	47.4	40.0	73.6	86.8	69.8	43.1	63.0	25.8	57.5
LLaVA-NeXT*-8B	0.6M	1.1M	82.5	64.8	41.3	47.4	60.8	47.7	38.0	72.1	85.3	69.4	44.2	59.9	26.2	56.9
LLaVA-NeXT*-Qwen2-7B	0.6M	1.1M	82.5	64.9	44.6	48.9	61.7	49.3	41.7	73.3	86.4	70.2	44.7	61.0	30.1	58.4
MMEvol-8B	0.6M	1.6M	**83.4**	65.0	47.8	50.1	62.3	50.0	42.3	73.9	88.8	**78.8**	46.4	62.6	32.3	60.3
MMEvol-Qwen2-7B	0.6M	1.6M	83.1	**65.5**	**55.8**	**51.6**	**64.1**	52.4	**47.2**	74.7	**89.1**	77.6	**47.7**	63.9	**41.8**	**62.7**

Table 2: Comparison with state-of-the-art methods on 13 visual-language benchmarks. Our models consistently improve LLaVA-NeXT under a head-to-head comparison, using the same prompts and the same base LLM, showing the effectiveness of enhanced pretraining data quality. "PT" denotes the pre-training data scale, "IT" denotes the instruction tuning data scale, and "*" denotes the baseline model trained on the seed dataset. We mark the best performance **bold** and the second-best underlined.

to the fully open-source SOTA model Cambrain-1 (Tong et al., 2024), our approach—despite utilizing seed data sampled from Cambrain-1's training set—achieves superior performance, with an average gain of 3.2 points. This result highlights the importance of instruction data quality over sheer data volume in improving the MLLM capabilities.

When compared to MiniCPM-v2.5 (Yao et al., 2024), another open-source SOTA model, our method—despite the considerable disparity in training data volume—demonstrates notable improvements across key benchmarks. Specifically, MMEvol-8 outperforms MiniCPM-v2.5 by 3.4 points on HallBench, 2.5 points on MIA, and 13.6 points on MMSInst, particularly excelling in instruction following, visual reasoning, and reducing visual hallucinations. These results reaffirm our findings from ablation studies and qualitative analyses, demonstrating that MMEvol-8 significantly enhances model generalization and robustness. Furthermore, by leveraging our high-quality instruction data and utilizing the leading large language model Qwen2, we successfully train a superior MLLM from scratch within one day using 4×8 A100 GPUs. This further validates that high-quality instruction data is more impactful than large-scale, low-quality datasets in optimizing model performance.

5 Conclusion

In this work, we propose an image-text instruction evolution framework and explore the techniques, insights, and benefits of Evol-Instruct for enhancing the quality and quantity of image-text instruction data. We employ three distinct evolution methods to increase the complexity and diversity of instruction data based on a limited seed dataset while utilizing instruction elimination to filter out harmful data. The data evolved through three rounds of evolution is used to train a new model, demonstrating state-of-the-art (SOTA) performance across a comprehensive set of benchmarks. Future directions include exploring integrating image generation models to synthesize new images and perform dual evolution of images and texts, aiming to train even more robust foundational models.

6 Limitation

Despite the effectiveness of **MMEvol**, certain limitations remain. Due to computational constraints, we performed instruction evolution on 163K samples, which represents approximately 12% of the available data in existing instruction datasets. Additionally, our experiments were conducted using an 8B-scale model, meaning that results could further improve with larger models and expanded training sets. Another limitation stems from the use of the OpenAI GPT-4o mini API in our evolution process. While GPT-4o mini provides high-quality instruction generation, its reliance on a proprietary model limits reproducibility. In future work, we plan to replace it with open-source alternatives, such as Qwen2VL, to ensure greater accessibility and transparency in instruction evolution. Furthermore, while **MMEvol** effectively improves instruction complexity and diversity, future extensions could explore hierarchical evolution mechanisms that adaptively adjust instruction complexity based on model progression. Expanding long-tail visual concepts and introducing multimodal active learning could further refine model training, helping MLLMs navigate real-world multimodal challenges with even greater accuracy and robustness.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jonas Belouadi, Anne Lauscher, and Steffen Eger. 2023. Automatikz: Text-guided synthesis of scientific vector graphics with tikz. *arXiv preprint arXiv:2310.00367*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. 2024. $vila^2$: Vila augmented vila. *arXiv preprint arXiv:2407.17453*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

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. *arXiv preprint arXiv:2404.12390*.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913.

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. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *EMNLP*, pages 7837–7851.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*, pages 5648–5656.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*, pages 235–251.

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

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*, 123:32–73.

Alexander Cong Li, Ellis Langham Brown, Alexei A Efros, and Deepak Pathak. 2023a. Internet explorer: Targeted representation learning on the open web. In *International Conference on Machine Learning*, pages 19385–19406. PMLR.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023b. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *ICML*, pages 19730–19742.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023e. Unisa: Unified generative framework for sentiment analysis. In *ACM MM*, pages 6132–6142.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. In *NeurIPS*.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. 2024c. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *arXiv preprint arXiv:2407.15838*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Run Luo, Yunshui Li, Longze Chen, Wanwei He, Ting-En Lin, Ziqiang Liu, Lei Zhang, Zikai Song, Xiaobo Xia, Tongliang Liu, et al. 2024. Deem: Diffusion models serve as the eyes of large language models for image perception. *arXiv preprint arXiv:2405.15232*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162.

Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. 2024a. Design2code: How far are we from automating front-end engineering? *arXiv preprint arXiv:2403.03163*.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024b. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *NeurIPS*.

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 *CVPR*, pages 8317–8326.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Zhengwei Tao, Ting-En Lin, Xiancai Chen, Hangyu Li, Yuchuan Wu, Yongbin Li, Zhi Jin, Fei Huang, Dacheng Tao, and Jingren Zhou. 2024. A survey on self-evolution of large language models. *arXiv preprint arXiv:2404.14387*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Wenxuan Wang, Quan Sun, Fan Zhang, Yepeng Tang, Jing Liu, and Xinlong Wang. 2024b. Diffusion feedback helps clip see better. *arXiv preprint arXiv:2407.20171*.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. 2023. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*.

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 *CVPR*, pages 9556–9567.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986.

Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, et al. 2024. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *arXiv preprint arXiv:2407.07053*.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. In *NeurIPS*.

A Curation Details of Seed Data

LLaVA-Instruct (Liu et al., 2024b) is a dataset of image-text instructions based on the COCO (Chen et al., 2015) data source and generated using the OpenAI ChatGPT API. The image-text instruction format in this dataset primarily includes three types: dialogue-based question-answering, global descriptions, and complex reasoning. ShareGPT4V (Chen et al., 2023), on the other hand, is a dataset constructed or rewritten using the OpenAI GPT-4V API, based on image-text pairs from SAM (Kirillov et al., 2023), COCO, and other sources to introduce richer details into captions. Both LLaVA-Instruct and ShareGPT4V significantly advance the development of MLLMs (Hu et al., 2022; Li et al., 2023e; Si et al., 2024b) and are widely used. We integrate samples from these two datasets containing the same image by concatenating the corresponding instruction data lists. For samples with global descriptions but no instruction data, we use the GPT-4o-mini API to supplement the missing instruction data, similar to LLaVA-Instruct, resulting in a combined dataset of 133K samples. To ensure the diversity of the seed data, we also include additional scientific chart data. Specifically, we sample 30K entries from Cambrain-1 (Tong et al., 2024), covering various types of image-text instructions such as code generation, chart interpretation, scientific question-answering, document understanding, and mathematical reasoning, ultimately forming a seed dataset of 163K image-text instructions.

B Implementation Details

After three rounds of evolution and filtering, we obtain 447K high-quality image-text instruction data with diversity and complexity. This data, combined with the ALLaVA instruction dataset, forms the 600K instruction data segment of the training recipe. To ensure a fair comparison with other methods, we combine the instruction data with other commonly used image-text data into the final training recipe, as shown in the table 3. Notably, we find that the DataEngine (Tong et al., 2024) data contains many harmful mismatched image-text pairs. We use OpenAI GPT-4o API to filter out harmful data and obtain 20K effective image-text instruction data. More details about training settings can be found in table 4.

C Additional Visualization Results

We perform a long-tail distribution visualization analysis to examine the impact of fine-grained perceptual evolution on the visual object domain within the instruction data. As shown in fig. 9, fine-grained perceptual evolution substantially improves the representation of long-tail visual objects, maximizing the extraction of usable visual information from images. This refinement enhances image-text alignment granularity, increases instruction diversity, improves model generalization, and reduces visual hallucinations. Additionally, we sample a specimen from SEED-163K and display its evolution process in fig. 10. In round 1, we perform fine-grained perceptual evolution, leading to instruction data with more precise details, including actions and attributes. In round 2, interaction evolution shifts instruction forms from general question answering to creative poetry generation, increasing the diversity of instruction formats. In round 3, cognitive reasoning evolution adds reasoning steps to the answers in the instruction data, enhancing its complexity. Through multiple rounds of instruction evolution, we improve the diversity and complexity of the seed data.

We plot the performance of the model at every 1k step across 9 evaluation datasets in fig. 11 to observe the learning trends during training. We can observe that the model learns OCR-related capabilities and mathematical reasoning abilities relatively smoothly, while general perception and cognitive skills exhibit more challenges. This may stem from conflicts arising from multi-source training tasks. A phased learning approach based on the difficulty of different tasks could be adopted to achieve better performance. We also present additional visualization results to demonstrate the capabilities of our model. As shown in fig. 12, our model trained on this data exhibits strong visual reasoning, instruction following, and fine-grained perception capabilities. Additionally, it identifies nuances in meme content, validating the effectiveness and efficiency of **MMEvol**.

To investigate the reliability of the rewrites produced by GPT-4-o-mini, we conducted a manual evaluation of the data before and after the evolution process. Specifically, we first extracted 30 images of various types from the seed data to ensure diversity, keeping 5 relevant question-answer pairs for each image. Subsequently, we carried out the corresponding evolution in three different directions,

Category	Sources	Size	Ratio
VQA	VQAV2 (Goyal et al., 2017)	83K	5.1%
Knowledge	OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022) VG (Krishna et al., 2017), GeoQA (Gao et al., 2023)	243K	14.9%
Reasoning	GQA (Hudson and Manning, 2019)	72K	4.4%
Grounding	RefCOCO (Kazemzadeh et al., 2014)	48K	2.9%
OCR	OCR-VQA (Mishra et al., 2019), TextVQA (Singh et al., 2019) AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022) DocVQA (Mathew et al., 2021), DVQA (Kafle et al., 2018) Synthdog-EN (Kim et al., 2021), Datikz (Belouadi et al., 2023) TabMWP (Lu et al., 2022b), ArxivQA (Li et al., 2024)	270K	16.5%
Instruct	**MMEvol**, ALLaVA (Chen et al., 2024a)	650K	39.8%
Language	ShareGPT, WizardLM (Xu et al., 2023)	183K	11.2%
Science/Code	Design2Code (Si et al., 2024a), MathVision (Wang et al., 2024a) Geo170k (Gao et al., 2023), ScienceQA (Lu et al., 2022a) Websight (Li et al., 2023a), Cambrain-Data-Engine (Tong et al., 2024)	85K	5.2%

Table 3: The mixture of training recipe datasets with corresponding categories and sources. We collect these public dataset form internet.

Hyperparameter	Ablation Stage 1	Ablation Stage 2	SOTA Stage 1	SOTA Stage 2
language model	LLaMA 3 8b	LLaMA 3 8b	LLaMA 3 8b Qwen 2 7b	LLaMA 3 8b Qwen 2 7b
global batch size	128	128	128	128
batch size	4	4	4	4
learning rate	1e-3	5e-5	1e-3	5e-5
lr schedule	cosine	cosine	cosine	cosine
lr warmup ratio	0.03	0.03	0.03	0.03
weight decay	0	0	0	0
epoch	1	1	1	1
optimizer	AdamW	AdamW	AdamW	AdamW
cost	4h	0.1h	4h	20h
dataset	LLaVA Pretrain	Seed-6K/Evol-6k	LLaVA Pretrain	Dataset Mixture

Table 4: The detailed training setup for **MMEvol** and the hyper-parameters across the training stages.

ultimately obtaining 450 evolved question-answer pairs, which were then subject to scoring and filtering. The results were distributed among five experts for manual evaluation of the accuracy of the model evolution and the scoring filter. The data is summarized in the table 5. From the table, it is evident that the average success rate of evolution using MLLM can reach 90%, while the accuracy of the scoring filter can achieve 94%, indicating the reliability of MMEovel. Additionally, we provide detailed scoring cases in fig. 14, highlighted in red.

D Additional Ablation Study

We conduct an ablation study starting with a 1K seed dataset, focusing on the ratios of the three evolutionary directions. The results in table 10 indicate that equal probability allocation among the three evolution strategies yields the highest average performance, highlighting their collective importance in enhancing the complexity and diversity of the evolved data. To further validate the effectiveness of our prompt design, which explicitly incorporates complexity and diversity metrics, we replace the evolved prompts with their simplest baseline versions and evaluate them on the 1K dataset. As shown in table 9, the inclusion of measurable complexity and diversity metrics significantly enhances the quality of instruction evolution. Furthermore, we conduct scalability experiments to validate the potential of **MMEvol**. As shown in table 12, we can observe that both the extended model and the evolved data scale can effectively enhance the capabilities of the MLLMs, and the larger the model, the greater the capability gain, which validates the potential of our method to generate high-quality data.

As shown in the table 6, we achieve better results by replacing the closed-source model with the

Data ID	Expert	Image Categories	FP-Evol (0-5)	I-Evol (0-5)	CR-Evol (0-5)	I-Elim (0-15)
0,1,3,4,5,6	0	LandMark, OCR, Human & Clothes, Traffic, Living room, Sport	5,4,4,5,5,4	5,4,3,4,5,4	5,3,4,5,4,4	15,13,13,14,13,14
7,8,9,10,11,12	1	Kitchen, Office supplies & Tools, Plants, Animal, Sport, LandMark	5,5,4,5,4,4	5,4,5,5,4,4	5,5,4,4,5,4	14,15,13,15,14,13
13,14,15,16,17,18	2	Foods, LandMark, OCR, Human & Clothes, Traffic, Sport	4,4,3,5,4,5	5,4,4,4,4,5	4,5,5,4,5,5	14,14,15,13,14,15
19,20,21,22,23,24	3	Foods, Sport, LandMark, Office supplies & Tools, Plants, Traffic	3,4,5,5,5,4	3,4,5,5,5,5	5,5,5,5,5,5	13,15,14,15,15,15
25,26,27,28,29,30	4	Animal, Sport, Traffic, Landmark, Sport, Office supplies & Tools	4,5,5,5,5,5	4,5,5,5,4,5	5,5,3,5,5,5	14,15,14,15,14,15
		Average Scores	89.3%	88.7%	92%	94.5%

Table 5: **Human Evaluation of Instruction Evolution.** The table shows the accuracy of manual evaluations by five experts on **MMEvol**'s evolution and filtering in three directions.

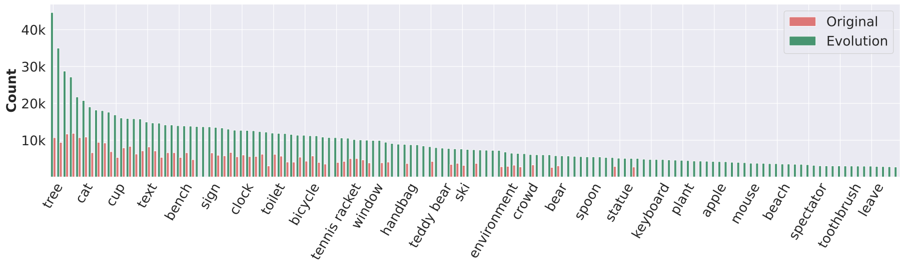


Figure 9: The long-tail distribution of 200 visual objects between seed and evolved data. **MMEvol** significantly improves the long-tail distribution of visual objects in the seed data, providing more fine-grained visual information, thereby boosting the model's generalization ability and robustness against hallucinations.

more advanced open-source Qwen2-VL, demonstrating the scalability and promise of our method. Furthermore, to fully showcase our technological contribution, we use previous SOTA methods (MIMIC-IT (Li et al., 2023b) and MMInstruct (Liu et al., 2024c)) to construct a dataset under the same seed data and API conditions, and conduct experiments. The results, as shown in the table 7, indicate that our method significantly outperforms previous methods which is consistent with our motivation. Additionally, our method is not restricted by data format or limited task forms, making it more user-friendly. We also perform robustness testing on the initial data of our method by selecting samples with a complexity score below 5 to construct a low-quality 1k seed dataset and conduct three rounds of evolution. As shown in the table 8, **MMEvol** is affected by low-quality initial seed samples, resulting in slightly lower outcomes compared to randomly sampled initial seed data. However, after multiple rounds of evolution, the gap narrows significantly, indicating **MMEvol**'s strong robustness and ability to quickly adapt and generate more complex and diverse high-quality samples.

E Complete Evolution Prompt Template

Due to the space limitations in the main text, we simplify the instruction evolution prompt template. We provide the complete detailed evolution templates as follows: the complete prefix-prompt template is shown in fig. 15, the fine-grained perception evolution prompt template is in fig. 16, the cognitive reasoning evolution prompt template is in fig. 17, the interaction evolution prompt template is in fig. 18, and the instruction elimination prompt template is in fig. 19.

VLM	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
GPT4o-mini (3K)	37.9	26.1	31.3	55.1	43.8	35.8	**53.2**	40.5
Qwen2VL-72B (3K)	**39.1**	**27.9**	**33.1**	**57.8**	**46.4**	**36.9**	46.9	**41.2**

Table 6: **Ablation study on VLM for instruction evolution.** If we utilize a more advanced open-source model such as Qwen2VL with 72 billion parameters, **MMEvol** could achieve further improvements.

Method	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
MIMIC-IT (3K)	32.1	24.3	26.4	47.6	41.9	31.5	34.5	34.1
MMInstruct (3K)	33.3	25.4	27.1	48.7	42.1	31.9	36.8	35.0
MMEvol (3K)	**37.9**	**26.1**	**31.3**	**55.1**	**43.8**	**35.8**	**53.2**	**40.5**

Table 7: **Ablation study on Method. MMEvol** demonstrates superior performance compared to the previous state-of-the-art methods, MIMIC-IT (Li et al., 2023b) and MMInstruct (Liu et al., 2024c), under the same conditions, validating the effectiveness and efficiency of iterative instructions with explicit optimization objectives.

Init	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
Low Score	36.5	25.4	29.9	54.8	43.1	35.0	52.6	39.7
Random	**37.9**	**26.1**	**31.3**	**55.1**	**43.8**	**35.8**	**53.2**	**40.5**

Table 8: **Ablation study on Initialization.** The quality of evolutionary data is influenced by the initial instructions, although the impact is relatively minor. Nevertheless, high-quality instructional data can still be generated through multiple iterations, demonstrating the robustness of our method.

FP-Evol	I-Evol	CR-Evol	I-Elim	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
✔	✔	✔	✔	34.7	25.7	29.9	54.1	42.1	35.5	49.8	38.8
✔	✔	✘	✔	35.7	25.9	30.3	54.8	42.9	35.2	51.2	39.4
✔	✘	✘	✔	36.5	25.4	30.8	55.0	43.6	35.4	52.4	39.9
✘	✘	✘	✔	**37.9**	**26.1**	**31.3**	**55.1**	**43.8**	**35.8**	**53.2**	**40.5**

Table 9: **Ablation study on prompt version.** UUtilizing carefully designed prompts significantly enhances data diversity and complexity, thereby improving the efficiency of the evolutionary process. The symbol ✔indicates that the prompt has been replaced with its baseline version during the evolutionary process.

FP-Evol	I-Evol	CR-Evol	I-Elim	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
2/3	1/6	1/6	✔	36.9	26.3	31.0	54.0	**44.8**	34.4	51.4	39.8
1/6	2/3	1/6	✔	34.3	25.4	29.2	53.2	43.5	**35.8**	52.6	39.2
1/6	1/6	2/3	✔	36.3	**26.7**	**32.5**	54.3	44.0	35.2	51.1	40.0
1/3	1/3	1/3	✔	**37.9**	26.1	31.3	**55.1**	43.8	**35.8**	**53.2**	**40.5**

Table 10: **Ablation study on instruction evolution ratio.** The highest average performance is achieved when the three evolutionary directions are assigned equal proportions, demonstrating their equal importance in enhancing the diversity and complexity of the evolved instruction data.

Skills	Sources	Skills	Sources
VQA	VQA^{v2} (Goyal et al., 2017)	General Knowledge	MME^{C} (Fu et al., 2023)
Knowledge Leakage	MMStar (Chen et al., 2024b)		MMMU (Yue et al., 2024)
Math Reasoning	MathVista^{M} (Lu et al., 2023)	Hallucination	POPE (Li et al., 2023d)
OCR Related	AI2D (Kembhavi et al., 2016)		HallBench (Guan et al., 2023)
Instruction Following	MIA (Qian et al., 2024)	Visual Reasoning	GQA (Hudson and Manning, 2019)
Visual Perception	BLINK (Fu et al., 2024), RWQA		MMSInst (Zhang et al., 2024)

Table 11: Benchmarks for evaluation with their sources and tested skills. The names are abbreviated due to space limitations. VQA^{V2}; GQA; VQA^{T}: TextVQA; MME^{C}: MME-Cognition; MathVista^{M}: MathVista-MINI; MMMU; AI2D; POPE; HallusionBench: HallBench; MIA; BLINK; RWQA: RealWorldQA; MMSInst: MM-Self-Instruct.

Seed

Question-seed: How many players are wearing orange jerseys?
Answer-seed: There are three players wearing orange jerseys.

Base Prompt

Question: How many individuals in the image are dressed in orange jerseys??
Answer: There are three individuals dressed in orange jerseys.

Fine-grained Perceptual Evolution

Question-R1: Observe the player wearing the **orange jersey with the number 8**. **How is their body positioned and what action are they engaged in**?
Answer-R1: The player in the orange jersey with the number 8 is in **mid-air, poised to take a shot, with both hands holding the ball**.

Interactive Evolution

Question-R2: Given the basketball game depicted in image, **create a poem that reflects the atmosphere of the moment captured**.
Answer-R2: On the vibrant court where spirits soar,\nThe game unfolds with passion's roar,\nPlayers leap and chase the score,\nIn pursuit of dreams and more.\nThe scoreboard's light flickers bright,\nEach play ignites the crowd's delight,\nIn this arena of youthful fight,\nWe see the glory of the night.

Cognitive Reasoning Evolution

Question-R3: Based on the information in the image, **which color team is likely to win?**
Answer-R3: **To determine** which team might win, **first**, we look at the scoreboard in the top-left corner of the image, which shows a score of 71:55. The team names on the scoreboard are not clearly visible. **Next**, we observe a blue player in the center of the image whose jersey displays the text 'MORABANC'. Given that the score of 71 corresponds to a team name that **can be identified as** 'MORABANC' from the jersey, and knowing the team's color is blue, **we infer that the blue team** (MORABANC) is likely to have an advantage in the current period.

Figure 10: **MMEvol** continuously enhances instruction data complexity and diversity over evol-instruct. The sample is from SEED-163K. We mark fine-grained visual information in red, new instructions form in green, and longer reasoning steps in blue. Additionally, we also provide the rewritten samples via a base prompt for comparison.

LLM	Data	MMStar	MathVista^{M}	MME^{C}	AI2D	HallBench	MMMU^{V}	RWQA	AVG.
Vicuna-7B	seed 3k	28.7	20.7	22.9	38.9	39.6	29.3	43.2	31.9
Vicuna-7B	evolved 3k	30.9	22.0	25.6	41.2	42.3	31.6	46.3	34.3
Vicuna-7B	evolved 6k	31.4	23.2	28.6	43.5	44.6	32.3	48.1	36.0
Vicuna-7B	evolved 9k	31.9	24.0	31.1	44.7	47.4	33.8	50.3	37.6
Vicuna-13B	evolved 9k	**34.6**	**26.1**	**34.5**	**50.6**	**52.3**	**36.1**	**54.5**	**41.3**

Table 12: **Ablation study on scalability.** The capabilities of model can be further enhanced with increasing data volume and model scale. Larger-scale models exhibit greater performance improvements when trained on our high-quality data.

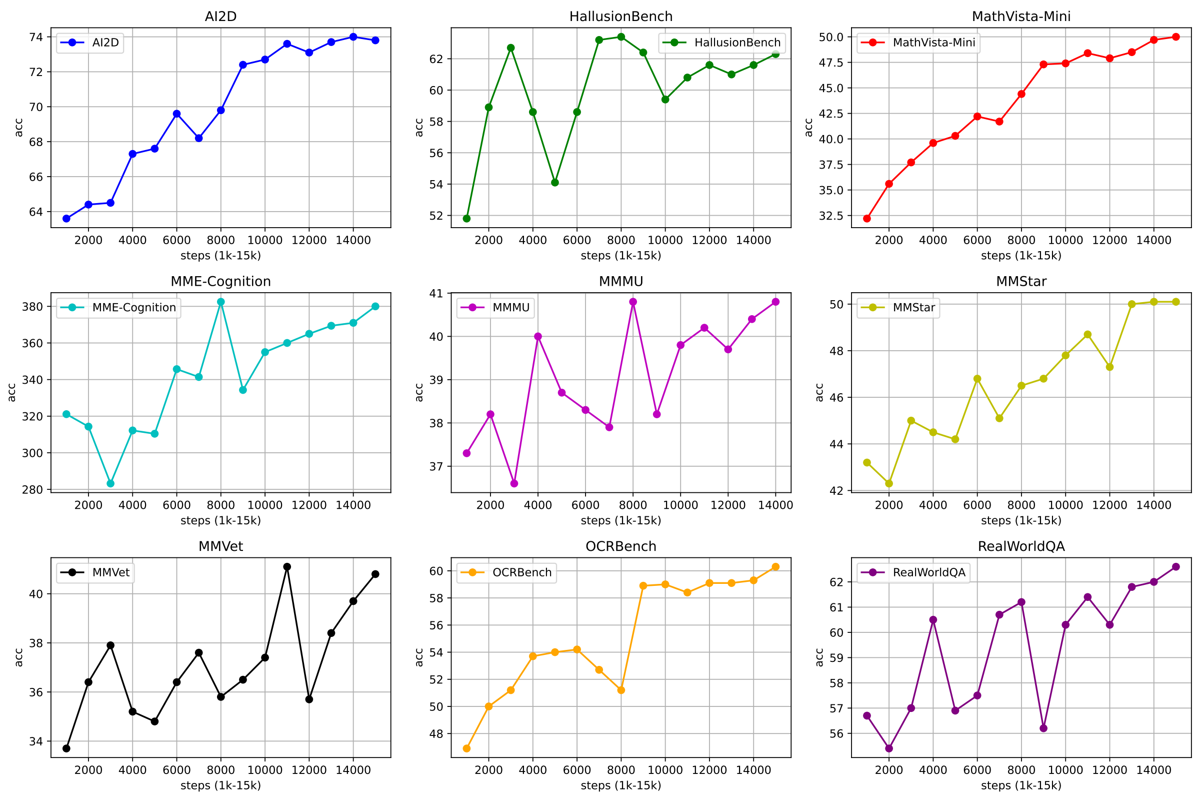


Figure 11: **Illustration of Performance Trends Across Various Benchmarks.** We plot the performance of the model at every 1k step across 9 evaluation datasets to observe the learning trends during training.

Figure 12: **Examples of image-text dialogue with our MMEvol-8B Model.** Training our model on the enhanced data not only enables it to handle handwritten mathematical reasoning, complex instruction following, and the understanding of emotional nuances, but also allows it to perceive fine-grained information within images, demonstrating the effectiveness of our approach.

Figure 13: **Some failed cases of image-text dialogue with our MMEvol-8B Model.** Our model generalizes to multiple visual reasoning scenes when training on the instruction data via our evolution strategy. However, there can still be errors in some scenarios, *e.g.*, hallucination.

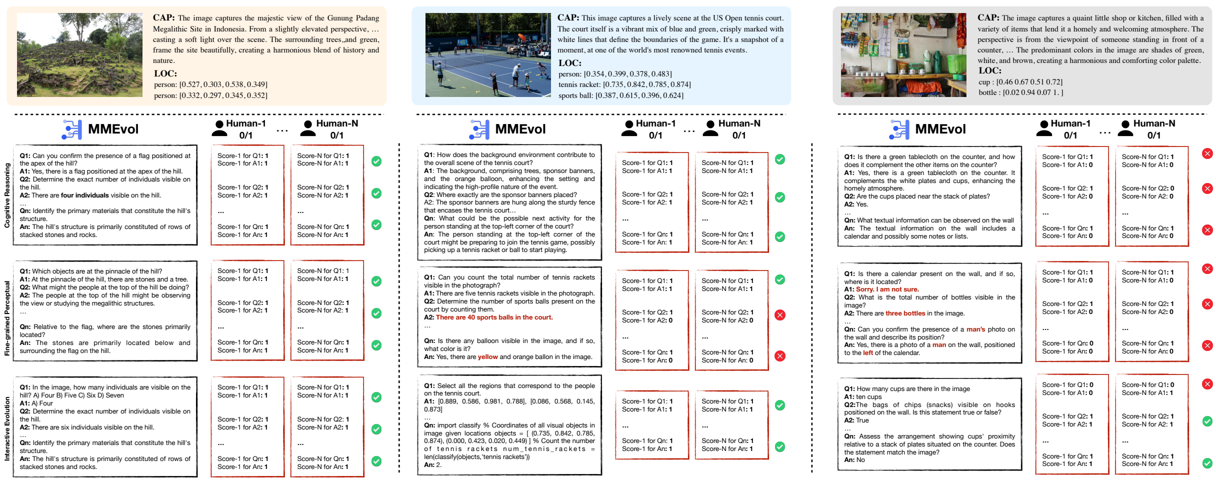


Figure 14: **Schematic diagram of the manual filtering process.** We hired N=5 experts to score (0/1) each question and answer. In the event that any question or answer receives a score of 0, the entire QA pair will be deemed invalid and discarded.

Context Type I: Caption

The image shows a modem living room with natural light streaming through a large window... A black couch against a gray wall, …, a glass coffee table that holds a white vase and a plant... The table rests on a beige rug, contrasting with the hardwood floor, adding warmth. The design suggests a comfortable and stylish living area.

Context Type II: Visual Object Locations

window : [0.2 0.23 0.57 0.4], couch : [0.17 0.43 0.83 0.79], vase : [0.5 0.51 0.58 0.72] …

Vision-Centered Multimodal Atomic Propositions & Permitted Vision-Centric Manipulations

1. Grounding Ability: Given a description of a visual object, output the coordinates of the visual object in the image and a natural language explanation.
2. Referencing Ability: Given the coordinates of a visual object, output the corresponding visual object description.
3. Calculating Ability: Ability to calculate the number, size, and other information of visual objects in the image and obtain the corresponding numbers.
4. OCR Ability: Recognize and generate textual representations of structured data in the image, such as numbers, text, codes, tables, etc.
5. Existence Ability: Given a description of a visual object, determine whether it exists in the image.

Permitted Vision-Centric Manipulations and Their Usage Descriptions

- Grounding_i(tgt)->bbx_i: The i-th grounding manipulation, that locates the object(s) specified by the target noun phrase `tgt` in the current image, and returns the resulting bounding box(es) as `bbx_i` where each box is represented by the top-left and bottom-right coordinates.
- Referring_i(bbx)->tgt_i: The i-th referencing manipulation, used to identify small and subtle objects in the image; it locates the current image using the box `bbx` defined by the top-left and bottom-right coordinates, zooms in the area by two times, and returns the resulting `tgt_i`.
- Calculate(tgt)->res_i: The i-th calculate manipulation, that calculates the formula specified by the target `tgt` in the current image, and returns the calculation result `res_i`.
- OCR_i(tgt)->txt_i: The i-th OCR manipulation, that recognizes the natural texts written on the target `tgt`, and returns the recognized texts `txt_i`.

Language-Centered Multimodal Atomic Propositions & Permitted Vision-Centric Manipulations

1. Relationship Description Ability: Understand and recognize relationships between different visual objects in the image, such as temporal, spatial, logical, etc.
2. Context Understanding Ability: Recognize and interpret complex scenes or situations in the image, such as asking about ongoing events, implied stories, unusual meaning, etc.
3. Behavior Prediction Ability: Predict possible subsequent actions based on the image content.
4. Knowledge Integration Ability: Integrate visual objects in the image with additional world knowledge, such as asking about background knowledge related to the objects.

Objective

…

Given Q&A

{

“**objects**” : [window, couch, vase, …]
“**skills**”: ["Grounding Ability", "Context Understanding Ability”, …]
"**format**" : "Complex reasoning",
“**question**”: "Which is closer to the window, couch or vase?"
"**steps**": [{"manipulation": "grounding_1(`window`)-> bbx_1"description": "Locate the window in the room and return its bounding box as `bbx_1`”.}, {"manipulation": "grounding_2(`couch`)-> bbx_2", "description": "Locate the couch in the room and return its bounding box as ` bbx_2}, {"manipulation": "grounding_3(`vase`)-> bbx_3", "description": "Locate the vase in the room and return its bounding box as `bbx_1`”.],
“**answer**”: First, based on the detected bounding boxes of the window [0.2, 0.23, 0.57, 0.4] and the couch [0.17, 0.43, 0.83, 0.79], it can be inferred that the couch is positioned below the window. Further, according to the location of the vase [0.5 0.51 0.58 0.72], it is located near the center of the room. Therefore, the couch is closer to the window.

}

Rewritten Q&A

Figure 15: **Complete prefix-prompt template of MMEvol.**

Objective of Fine-grained Perceptual Evolution
I want you act as a Q&A Creator. Your objective is to draw inspiration from the given Q&A to create a brand new created Q&A. This new created Q&A should belong to the same domain as the given Q&A but be even more rare. The difficulty level of the created Q&A should be similar to that of the given Q&A. Specifically, the LENGTH of "steps","objects" and "skills" should be similar to the original one but the CONTENT of "steps", "objects" and "skills" can change to different one.
Prioritize questions with definite answers. If a question can be resolved with only a few solving steps, it can be reformulated to explicitly request additional solving steps. It is essential to avoid making the #Rewritten Q&A overly verbose.

Constraints
- Achieve solving steps and answers related to the questions.
- Ensure all generated data is consistent with the image content.
- Double-check provided descriptions against the image content.
- Do not generate new location coordinates; use the given coordinates.
- Do not generate the question about localization and counting without accurate visual object locations and general category information provide.

Example
...

Given Q&A
…

Rewritten Q&A
...

Figure 16: **Complete fine-grained perceptual evolution prompt template.**

Objective of Cognitive Reasoning Evolution
I want you act as a Q&A Rewriter. Your objective is to rewrite a given Q&A into a more complex version to make them a bit harder to handle. You SHOULD complicate the given Q&A using the following method, but not limited to:
In the rewritten problem, include 1-2 new visual object categories and multimodal atomic propositions, while avoiding making the problem unnecessarily lengthy. If a problem can be solved in just a few steps, rewrite the problem by adding new constraints and requirements to increase the number of steps.

Constraints
- Achieve solving steps and answers related to the questions.
- Ensure all generated data is consistent with the image content.
- Double-check provided descriptions against the image content.
- Do not generate new location coordinates; use the given coordinates.
- Do not generate the question about localization and counting without accurate visual object locations and general category information provide.

Example
...

Given Q&A
…

Rewritten Q&A
...

Figure 17: **Complete cognitive reasoning evolution prompt template.**

Objective of Interactive Evolution
I want you act as a Q&A Rewriter. Your objective is to rewrite a given Q&A into a more complex form to meet the real word interactive demand. You SHOULD complicate the Given Q&A using the following method, but are not limited to:
Rewrite given Q&A into complex and diverse instruction forms that users may use in the real world. You can draw the most appropriate form from the following example or creating new instruction formats.

Prioritize questions with definite answers.
If a question can be resolved with only a few solving steps, it can be reformulated to explicitly request additional solving steps. It is essential to avoid making the #Rewritten Q&A# overly verbose.

Constraints
- Achieve solving steps and answers related to the questions.
- Ensure all generated data is consistent with the image content.
- Double-check provided descriptions against the image content.
- Do not generate new location coordinates; use the given coordinates.
- Do not generate the question about localization and counting if there are no accurate visual object locations and general category information provide.
- Ensure the image position in the generated Q&A&S is consistent with the given Q&A&S, and that they all belong to the same sub-image.

Example
...

Given Q&A
...

Rewritten Q&A
...

Figure 18: **Complete interactive evolution prompt template.**

Instruction Elimination
Evaluate the difficulty and complexity of each rewritten samples compared to its original ones. Determine if the rewritten sample has improved in quality via "yes/no".
Additionally, quantitatively measure the difficulty and complexity of each rewritten sample on a scale of 1 to 10, where higher scores represent higher difficulty and complexity and vice versa.
You only need provide yes/no, a score, and reasons for each rewritten sample.

{“improved”: “yes/no”, “score”: 1-10, “reason”: the reason for the improvement and score},

Evaluation Criteria
- Length: Longer Q&As generally have more detail and thus are considered more complex.
- Semantic Complexity: Use of more sophisticated language or concepts.
- Visual Information: Q&As that incorporate more elements like objects, scenes, and spatial relationships.
- Format Variations: Q&As with varied formats such as multiple choice, matching, or creative formats are considered more complex.
- Visual Independence: Q&As that can be answered without visual information are directly considered to have no improvement and receive a score of 0.
Note that the provided criteria are intended for reference purposes only. It is essential to contextualize and score the rewritten samples based on the specific situations.

In-context QA samples with different difficulties (1-10)
...

Figure 19: **Complete instruction elimination prompt template.**

Cognitive Reasoning Evolution

conversations_v_0

Q: What is the name displayed on the clock tower?
A: The name 'BRITANNIA' is displayed on the clock tower.

Q: Where is the clock located on the building?
A: The clock is located on the upper part of the clock tower, with coordinates approximately [0.488, 0.259, 0.553, 0.348].

Q: How does the sky contribute to the overall atmosphere of the image?
A: The light blue sky dotted with fluffy white clouds contributes to a bright and harmonious atmosphere in the image. It complements the blend of old and new architectural styles, unifying the scene under a shared expanse.

conversations_v_1

Q: What name is prominently displayed on the clock tower, and how does it relate to the architectural features of the clock and building? Let's think step by step.
A: The name 'BRITANNIA' is displayed on the clock tower. This name contributes to the identity of the structure, reflecting its historical significance in the urban landscape. The brick building, characterized by its warm brown hue and arched windows, adds a sense of nostalgia and tradition that contrasts with the modern skyscrapers in the background. Together, these elements illustrate a conversation between past and present in the architectural aspects of the scene.

Q: Given the positioning of the clock on the building, how does this contribute to the overall function and visual prominence of the tower in the urban scene? Let's think step by step.
A: The clock is positioned at the upper part of the clock tower, with bounding coordinates approximately [0.488, 0.259, 0.553, 0.348]. This elevated placement enhances its visibility, allowing it to serve as a key point of orientation within the city. As a prominent feature, the clock not only fulfills its practical function of telling time but also acts as a striking landmark, drawing attention within the contrasting backdrop of the sleek modern skyscrapers nearby.

Q: In what ways does the sky contribute to the overall aesthetic and functional significance of the clock tower within this urban scene? Let's think step by step.
A: The light blue sky adorned with fluffy white clouds infuses the urban scene with brightness and tranquility. This serene backdrop not only enhances the aesthetic appeal of the clock tower but also serves to focus attention on it as a significant landmark within the skyline. The harmonious atmosphere created by the sky juxtaposes beautifully with the architectural elements of the buildings, reinforcing the clock tower's role as a central feature in the city's identity.

Figure 20: **Data case of cognitive reasoning evolution.**

Fine-grained Perceptual Evolution

conversations_v_0

Q: What material is prominently featured in the kitchen?
A: Wood is the prominently featured material in the kitchen, as seen in the table, chairs, cabinets, and floors.

Q: What is the condition of the kitchen in the image?
A: The kitchen appears to be immaculate, clean, and well-maintained.

Q: What kind of countertops does the kitchen have?
A: The kitchen has dark countertops that contrast with the wood elements.

Q: Is there an island or a small table in the kitchen?
A: There is an island and a small table present in the home kitchen, providing additional workspace and seating options.

Q: What type of flooring material is used in this kitchen?
A: The flooring material used in this kitchen is beautiful hardwood floors.

conversations_v_1

Q: What types of appliances are integrated into the kitchen?
A: The kitchen is equipped with modern appliances including a gas range, refrigerator, dishwasher, microwave, and coffee maker.

Q: What type of lighting fixture is predominately used in the kitchen?
A: The kitchen features pendant lights above the island, providing style and bright illumination.

Q: Is there any decorative shelf visible in the kitchen?
A: Yes, there is a wooden shelf adorned with a plant and vase, adding a touch of nature to the kitchen.

Q: What type of view does the kitchen window offer?
A: The kitchen window offers a picturesque view of a rock wall and lush greenery outside.

Q: What does the dining setup in the kitchen suggest about its use?
A: The dining setup with a table, four chairs, and a bench suggests that the kitchen can comfortably accommodate family meals and gatherings.

Figure 21: **Data case of fine-grained perceptual evolution.**

Interactive Evolution

conversations_v_0

Q: How many zebras are visible in the image?
A: There are twelve zebras visible in the picture. Their bounding box coordinates are as follows: [0.216, 0.366, 0.834, 0.946], [0.681, 0.517, 1.000, 0.937], [0.105, 0.293, 0.451, 0.578], [0.440, 0.317, 0.683, 0.551], [0.647, 0.038, 0.999, 0.337], [0.657, 0.032, 0.877, 0.156], [0.519, 0.048, 0.683, 0.321], [0.236, 0.069, 0.614, 0.321], [0.048, 0.022, 0.365, 0.235], [0.229, 0.418, 0.391, 0.557], [0.000, 0.358, 0.234, 0.950], and [0.841, 0.276, 1.000, 0.494].

Q: Are there any zebras grazing on grass in the image?
A: Yes, there are zebras grazing on grass in the image, as evidenced by some of the zebras lowering their heads to the grass.

Q: How many zebras are in the foreground?
A: There are three zebras in the foreground, represented by the larger and more prominently located zebras.

Q: What is the relationship between the zebras and their habitat in the image?
A: The zebras are seen grazing and standing alert in their natural habitat, which is a dry grass field. This type of habitat provides food and a space where they can stay in groups for safety.

conversations_v_1

Q: Select the accurate number of zebras visible in the image from the options: A) 10 B) 12 C) 15 D) 20.
A: B

Q: Between the zebras located in the foreground and background, which group has a higher count?
A: There are three zebras in the foreground, represented by the larger and more prominently located zebras.

Q: Provide two rationales for how the zebras interact with their habitat depicted in the image.
A: 1. The zebras are seen grazing on the dry grass, utilizing it as a food source. 2. Their presence in the grassland allows them to maintain social interactions and observe their surroundings for predators.

Figure 22: **Data case of interactive evolution.**