A. Additional Experiments and Discussions

A.1. Relation to Magpie

Magpie is a framework for efficient LLM training data synthesis that leverages a similar "hooking" method like ours. In LLM scenarios, Magpie divides the input to an LLM into three parts: the pre-query template, the query, and the post-query template. For Llama-3-8B-Instruct, an input can be "`<|start_header_id|>user <|end_header_id|>Hi! <|start_header_id|> assistant<|end_header_id|>`". They feed only the pre-query template (blue part) into Llama-3-8B-Instruct and extract potential instruction output, leading to 300K high-quality and diverse instances, which could be further extended easily. They use their collected data to fine-tune Llama-3 and achieve remarkable advantages against six other state-of-the-art open-source instruction tuning datasets.

We gain our inspiration from Magpie and extend the method to multimodal scenarios. Consequently, we explore the feasibility of this idea to synthesize multimodal data in depth and propose a novel method, **Oasis**, which could lead to huge improvements in MLLMs. Compared to Magpie, a primary difference in our method is that we include the image as an additional input, which allows the MLLM to generate instruction based not only on its internal knowledge but also on the visual information. This attribute enables the image domain to control the synthesized instruction domain, making our method more versatile and applicable to a broader range of tasks. It is also worth noting that we handcraft all-around quality control means specifically for multimodal data, which effectively ensures the quality of the synthesized data and paves the way for the community to explore better multimodal data synthesis.

A.2. Cases of Oasis-500k

We provide more cases of **Oasis**-500k data in Fig. 7. It can be observed that the generated instruction encompasses a wide range of tasks and domains, including OCR, object recognition, scenario understanding, commonsense knowledge, etc. Thanks to the 'hooking' method, the instruction is diverse, creative, and enlightening, which is beneficial for the multimodal model to extend its generalization ability.

A.3. Instruction Quality Control Details

In Step 3 of our method, a comprehensive quality control process is conducted to ensure the quality of the synthesized data. In detail, we evaluate the solvability, hallucination, clarity, and nonsense of the instruction and filter out 50% of the data. Here we provide the detailed filtering criteria for each dimension. Each dimension of instruction is scored on a scale of 1 to 5, with 1 being the worst and 5 being the best. For hallucination and nonsense, we only retain the data with a score of 5, since the existence of any hallucination or nonsense could lead to misleading training, harming the model's performance and generalization ability. For solvability and clarity, only the data that satisfies each score being greater than or equal to 3, and the sum of the scores being greater than or equal to 7 will be retained. This standard is set as a balance of filterability, synthesis efficiency, and data diversity.

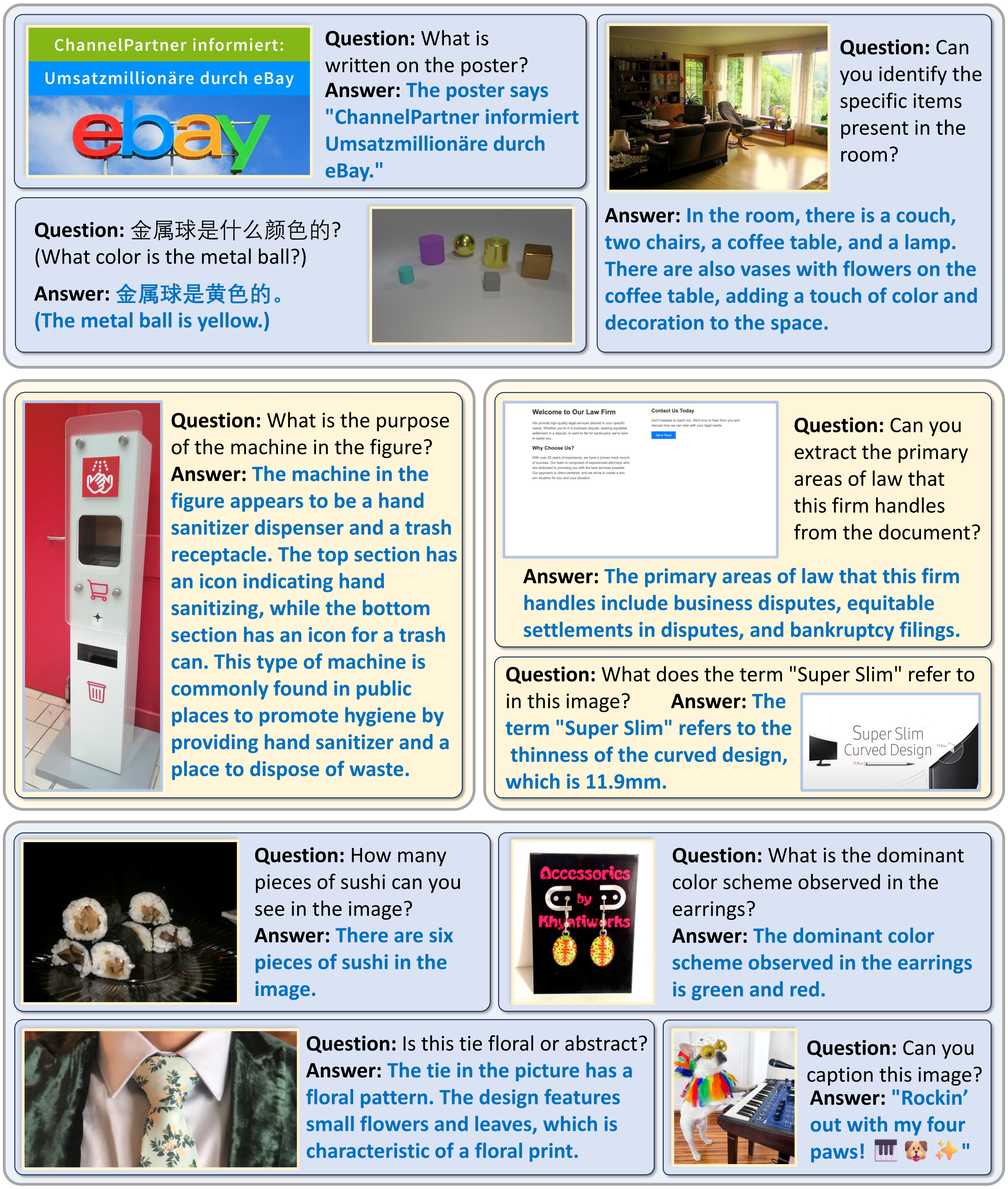
Distribution of instruction quality scores. Figure 8 here illustrates the range and distribution of quality scores assigned to instructions. The results show that the majority of instructions are rated highly in terms of hallucination and nonsense, which can be attributed to the strength of MLLM. In comparison, the solvability and clarity scores are more evenly distributed, which leads to a sufficient filtering mechanism.

A.4. More data analysis

Oasis data has large type-token ratios. We calculate the type-token ratio (TTR) of the instruction and response data. The TTR is defined as the ratio of the number of unique words to the total number of words in the dataset. As shown in Tab. 6, the TTR of **Oasis** is significantly higher than that of LLaVA-NeXT, especially in the instruction data. This indicates that **Oasis** data is more lexically diverse and covers a wider range of topics, which can help improve the generalization ability.

A.5. Application on medical area

We validate **Oasis** on medical benchmarks in Tab. 7. We sample 15k images from the MedTrinity-25M dataset and create 2k medical training data with **Oasis**. We SFT the LLaVA-NeXT baseline with 4k sampled LLaVA data and 2k LLaVA data + 2k synthesized medical data, respectively. The table below shows great performance improvements across 3 medical benchmarks with our data.



Question: What is written on the poster?
Answer: The poster says "ChannelPartner informiert Umsatzmillionäre durch eBay."

Question: Can you identify the specific items present in the room?
Answer: In the room, there is a couch, two chairs, a coffee table, and a lamp. There are also vases with flowers on the coffee table, adding a touch of color and decoration to the space.

Question: 金属球是什么颜色的? (What color is the metal ball?)
Answer: 金属球是黄色的。(The metal ball is yellow.)

Question: What is the purpose of the machine in the figure?
Answer: The machine in the figure appears to be a hand sanitizer dispenser and a trash receptacle. The top section has an icon indicating hand sanitizing, while the bottom section has an icon for a trash can. This type of machine is commonly found in public places to promote hygiene by providing hand sanitizer and a place to dispose of waste.

Welcome to Our Law Firm

We provide high-quality legal services tailored to your specific needs. Whether you're in a business dispute, seeking equitable settlement in a dispute, or want to file for bankruptcy, we're here to assist you.

Why Choose Us?

With over 25 years of experience, we have a proven track record of success. Our team is composed of experienced attorneys who are dedicated to providing you with the best services possible. Our approach is client-centered, and we strive to create a win-win situation for you and your situation.

Contact Us Today

Don't hesitate to reach out. We'd love to hear from you and discuss how we can help with your legal needs.

Get in Touch

Question: Can you extract the primary areas of law that this firm handles from the document?
Answer: The primary areas of law that this firm handles include business disputes, equitable settlements in disputes, and bankruptcy filings.

Question: What does the term "Super Slim" refer to in this image? **Answer:** The term "Super Slim" refers to the thinness of the curved design, which is 11.9mm.

Question: How many pieces of sushi can you see in the image?
Answer: There are six pieces of sushi in the image.

Question: What is the dominant color scheme observed in the earrings?
Answer: The dominant color scheme observed in the earrings is green and red.

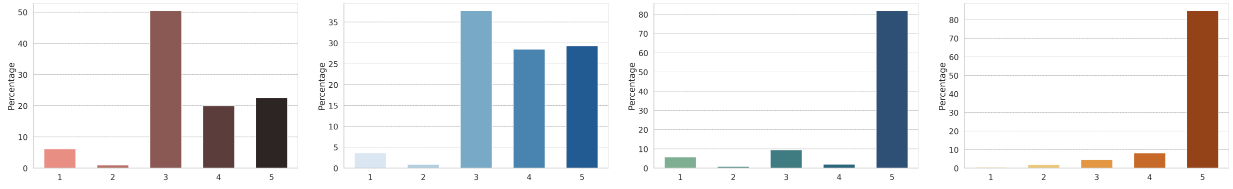
Question: Is this tie floral or abstract?
Answer: The tie in the picture has a floral pattern. The design features small flowers and leaves, which is characteristic of a floral print.

Question: Can you caption this image?
Answer: "Rockin' out with my four paws! 🎹 🐶 ✨"

Figure 7. **Oasis data cases.** This figure shows several cases of **Oasis** data. It can be observed that the data synthesized by **Oasis** is diverse and creative, covering a wide range of tasks and domains.

B. Prompts for Data Filtering

Oasis contains 2 steps of data filtering: data categorization and quality control. We carefully design the filtering logic to ensure the validity and quality of the data. The efficacy of the filtering process is crucial for the success of our method. Therefore, we handcraft specific prompts for each filtering step to make sure the data is correctly categorized and rated in multiple dimensions. The following are the prompts used in the data filtering process.



(a) Solvability Score (b) Clarity Score (c) Hallucination Score (d) Nonsense Score

Figure 8. **Distribution of instruction quality scores.** We use MLLM to evaluate the solvability, clarity, and hallucination scores, and LLM to evaluate the nonsense score. The majority of instructions contain no hallucination and nonsense, while the solvability and clarity attributes are more evenly distributed.

Table 6. **Type-token ratio statistics**. Type-token ratio (TTR) analysis of LLaVA-NeXT SFT data and **Oasis**-500k.

Metric	LLaVA-NeXT	**Oasis**-500k
Instruction TTR	0.0018	0.0124
Response TTR	0.0064	0.0086

Table 7. **Oasis in medical**. Medical data generated by **Oasis** leads to consistent improvements across 3 medical benchmarks.

Data	PathVQA	VQA-RAD	SLAKE(en)
4k LLaVA	34.0	47.5	49.0
2k LLaVA + 2k Oasis	**41.2**	**48.1**	**56.7**

B.1. Data Categorization

The data generated by the first step of **Oasis** can be generally categorized into 2 types: image caption and instruction, and only the latter should be retained. Because of the nature of our method, it is observed that the instruction can often hide in large tracts of text, and it is often followed by an answer, which is undesirable. Moreover, the instruction can be in various forms, such as interrogative sentences, imperative sentences, multiple-choice questions, etc. In order to achieve better categorization accuracy, we leverage few-shot examples and design the following prompts.

```
You will be given a text regarding an image. Your task is to determine whether the text
    contains any instructions. If it contains instructions, extract one instruction. You
    should extract the instruction, as well as any relevant contextual information that aids
     in understanding the instruction.

NOTE:
1. The instruction may take the form of an interrogative sentence, an imperative sentence, a
     multiple-choice question, or other similar structures. Please identify carefully!
2. Extract ONLY the original instruction, WITHOUT extracting any answers.
3. If the instruction is a multiple-choice question, you should extract the question and the
     options.
4. If there are multiple instructions, you should extract only one instruction.

You MUST answer with the following format:
Instruction: [an instruction]

If it doesn't contain any instructions, output 'NO_INST'.

----- Example 1:
Text:
1. Answer the following questions based on the text:\n\n    a. Who increased the number of
    insurgents in the valley? \n\n    b. When did Singh come to power? What act did he
    implement?\n\n    c. What is the purpose of the SC-ST Act?

Answer:
Instruction: Who increased the number of insurgents in the valley?

----- Example 2:
Text:
There is an animal behind the fence who is holding a bottle.
```

```

Answer:
NO_INST

----- Example 3:
Text:
In this problem, we have an elephant image that includes several lines and curves.\n\nWe want to transform this image into another animal using the least number of changes.\n\nPlease provide some suggestions on how to achieve this transformation with minimal effort.

Answer:
Instruction: We want to transform this image into another animal using the least number of changes.\n\nPlease provide some suggestions on how to achieve this transformation with minimal effort.

----- Example 4:
Text:
Could you please summarize the mission statement of the company and the benefits it promises to its customers in 30 seconds or less?\n The mission of our company is to provide innovative tech solutions for all your needs. We prioritize security and privacy for our users and are committed to excellence. With us by their side, customers can expect a simplified tech journey that feels more defined.

Answer:
Instruction: Could you please summarize the mission statement of the company and the benefits it promises to its customers in 30 seconds or less?

----- Example 5:
Text:
I would like to make a real estate agency website using HTML, CSS, and JavaScript.

Answer:
Instruction: I would like to make a real estate agency website using HTML, CSS, and JavaScript.

----- Example 6:
Text:
The scene has a window on the top left, a fire hydrant on the bottom right, and two signs in the middle right.

Answer:
NO_INST
----- End of Example

[Begin of Text]
{text}
[End of Text]
```

B.2. Instruction Quality Control

The quality control stage evaluates the comprehensive quality of the instructions, including solvability, hallucination, clarity, and nonsense. This step directly determines the representation ability of the data and thus the performance of the model, so the quality control process should be effective and rigorous. We notice that models have a strong tendency to score a 5 with plain prompt, which could lead to biased and insufficient filtering. Therefore, we list the specific scoring criteria for each score in the final prompt.

Prompt for solvability.

```
Your task is to evaluate the solvability of a query to an image. The solvability can be quantitatively evaluated on a scale of 1 to 5, based on the presence of sufficient information within the image to formulate a complete answer.

Here are the criteria:

Score 1 (Very Low Solvability): The image contains minimal or no relevant information related to the question, making it nearly impossible to derive a meaningful answer.

Score 2 (Low Solvability): The image provides some information, but key elements are missing, resulting in significant uncertainty.

Score 3 (Moderate Solvability): The image contains a reasonable amount of information that may lead to an answer, but ambiguities or lack of clarity hinder definitive conclusions.

Score 4 (High Solvability): The image offers substantial information that strongly supports answering the question, with only minor uncertainties remaining.

Score 5 (Very High Solvability): The image is rich in detail and clarity, providing all necessary information to answer the question comprehensively.

Please rate the query on a scale of 1 to 5. Use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate your evaluation score in the key 'Score'.

[Query]
{query}
```

Prompt for hallucination.

```
Your task is to evaluate whether a query to an image contains hallucination content. The determination of whether a question related to an image contains hallucinations can be assessed on a scale of 1 to 5. This scale evaluates the alignment between the question's content and the actual content of the image, identifying discrepancies that indicate hallucinations.

Here are the criteria:

Score 1 (Severe Hallucination): The question bears little to no relation to the image content, filled with substantial errors or completely unrelated information. The discrepancies are so pronounced that they render the question fundamentally flawed in context to the image.

Score 2 (Significant Hallucination): The question diverges considerably from the image, containing multiple erroneous statements or irrelevant details. The inaccuracies are significant enough that they compromise the integrity of the inquiry.

Score 3 (Moderate Hallucination): The question and image content have notable inconsistencies, with several inaccuracies present. While some relevant information is shared, the question includes errors that could lead to misleading conclusions.

Score 4 (Minor Hallucination): The question is largely consistent with the image, but there are minor discrepancies or inaccuracies that do not significantly alter the overall interpretation. These could include slight misinterpretations of color or detail that do not affect the main subject.

Score 5 (No Hallucination): The question aligns perfectly with the image content, containing no errors or irrelevant information. All aspects of the inquiry are directly supported by clear and accurate details within the image.
```

```

Please rate the query on a scale of 1 to 5. Use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate your evaluation score in the key 'Score'.

[Query]
{query}
```

Prompt for clarity.

```
Your task is to evaluate the clarity of a query to an image. The clarity of a question derived from an image can be evaluated on a scale of 1 to 5, reflecting how precisely the question conveys its intent and whether it allows for a definitive answer.

Here are the criteria:

Score 1 (Very Unclear): The question is exceedingly vague and unclear, with multiple interpretations possible. It fails to convey a coherent intent, resulting in uncertainty and an inability to arrive at a definitive answer.

Score 2 (Unclear): The question is largely ambiguous, making it difficult to discern its exact intent. The vagueness significantly hinders the ability to provide a clear answer, leading to potential misinterpretations and disagreements.

Score 3 (Moderately Clear): The question exhibits noticeable vagueness that may cause some confusion. While there are identifiable elements, the lack of precision can lead to varying interpretations and uncertainty in answering.

Score 4 (Clear): The question is generally clear but may contain minor ambiguities that could lead to slight misinterpretations. However, the overall intent remains understandable, allowing for a reasonably definitive answer.

Score 5 (Very Clear): The question is exceptionally clear, leaving no room for ambiguity. It conveys its intent explicitly, and the required answer is straightforward and unambiguous, making it easy to interpret.

Please rate the query on a scale of 1 to 5. Use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate your evaluation score in the key 'Score'.

[Query]
{query}
```

Prompt for nonsense.

```
Your task is to evaluate whether a query to an image contains nonsense. The presence of nonsense in a question related to an image can be assessed on a scale of 1 to 5.

Here are the criteria:

Score 1 (Severe Nonsense): The question is completely nonsensical, filled with severe grammatical issues, strange characters, or illogical phrases that render it unintelligible. It fails to convey any meaningful intent.

Score 2 (Significant Nonsense): The question is largely incoherent, containing multiple grammatical errors or strange characters that obstruct its meaning. Understanding the question is challenging and may lead to misinterpretations.

Score 3 (Moderate Nonsense): The question exhibits noticeable issues with clarity, such as awkward constructions or vague expressions. While some meaning is still discernible, these factors may lead to confusion.

Score 4 (Minimal Nonsense): The question is generally clear but may contain minor grammatical errors or awkward phrasing that slightly detract from its coherence. These issues do not significantly impede understanding.

Score 5 (No Nonsense): The question is coherent, grammatically correct, and free from any strange characters or phrases. It conveys its intent clearly and logically, allowing for a straightforward understanding.

Please rate the query on a scale of 1 to 5. Use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate your evaluation score in the key 'Score'.

[Query]
{query}
```