Express 💬 What You See 👀: Can Multimodal LLMs Decode Visual Ciphers with Intuitive Semiosis Comprehension?

Jiayi Kuang[1,*], Yinghui Li[2,*], Chen Wang[3], Haohao Luo[1], Ying Shen[1, 5,†], Wenhao Jiang[4]

[1]Sun Yat-sen University, [2]Tsinghua University, [3]University of Pennsylvania,
[4]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ),
[5]Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology

[*]These authors contributed equally. [†]Corresponding author: sheny76@mail.sysu.edu.cn

Abstract

Bridging the gap between visual and language remains a pivotal challenge for the multimodal community. Traditional VQA benchmarks encounter a modality gap and over-reliance on language priors, whereas human cognition excels at intuitive semiosis, associating abstract visual symbols to linguistic semantics. Inspired by this neurocognitive mechanism, we focus on emojis, the visual cipher conveying abstract textual semantics. Specifically, we propose a novel task of generating abstract linguistics from emoji sequence images, where such reasoning underpins critical applications in cryptography, thus challenging MLLMs' reasoning of decoding complex semantics of visual ciphers. We introduce **eWe-bench** (**E**xpress **W**hat you Se**E**), assessing MLLMs' capability of intuitive semiosis like humans. Our data construction framework ensures high visual sensitivity and data quality, which can be extended to future data enhancement. Evaluation results on advanced MLLMs highlight critical deficiencies in visual intuitive symbolic reasoning. We believe our interesting insights for advancing visual semiosis in MLLMs will pave the way for cryptographic analysis and high-level intuitive cognition intelligence of MLLMs [1].

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable progress in recent years (Liu et al., 2022; Dong et al., 2023; Xu et al., 2024; Yin et al., 2023; Wu et al., 2023; Cui et al., 2024; Kuang et al., 2024; Li et al., 2025a), particularly in their ability to integrate visual and linguistic information for more natural human-computer interaction (Li et al., 2022c; Zhang et al., 2024a; Luo et al., 2024; Li et al., 2024e). Despite these advances, a central challenge remains: ***how can MLLMs truly bridge the gap between vision and language to emulate human-like perception?*** (Koh et al., 2024; Peng et al., 2023; Wang et al., 2024). Traditional Visual Question Answering (VQA) based MLLM benchmarks have driven progress by evaluating from various dimensions (Li et al., 2024b, 2023a) and different domains (Lu et al., 2024b; Yue et al., 2024; Zhang et al., 2024b). However, their inherent designs of decoupling images and questions have fragmented visual-language representations, which introduces critical limitations in various language scenarios: (1) over-reliance on language priors instead of visual understanding, and (2) question-guided reasoning that masks spontaneous visual intuition (Peng et al., 2023; Wang et al., 2024).

Human cognition presents a different paradigm. Neuroscientific studies reveal that our brains reflect intuitive semiosis capability, which spontaneously maps abstract visual symbols to semantic representations through rapid thalamocortical interactions (Schulze Buschoff et al., 2025; Islam and Bouwman, 2016; Hobson and Pace-Schott, 2002), with symbol recognition occurring 200-300 milliseconds faster than object categorization. This ***visual symbol intuitive semiosis*** enables instantaneous understanding of abstract symbols that convey complex meanings through shared cultural codes, including traffic signs, mathematical notations, and emojis (Du et al., 2023). Such visual symbol reasoning underpins critical applications in cryptography and information analysis (Ateniese et al., 1996; Moulin and O'Sullivan, 2003), where successful decoding requires direct symbol-to-semantic mapping without explanatory prompts.

Motivated by this mechanism, we propose a transformative benchmark to evaluate MLLMs' ***visual intuitive semiosis comprehension*** in multilingual scenarios. Departing from the VQA paradigm that uses textual questions as cognitive scaffolds, our approach employs emojis as ***visual ciphers***, challenging models to decode abstract visual sym-

[1] All our data are available in eWe-bench.

Findings of the Association for Computational Linguistics: ACL 2025, pages 12743–12774
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

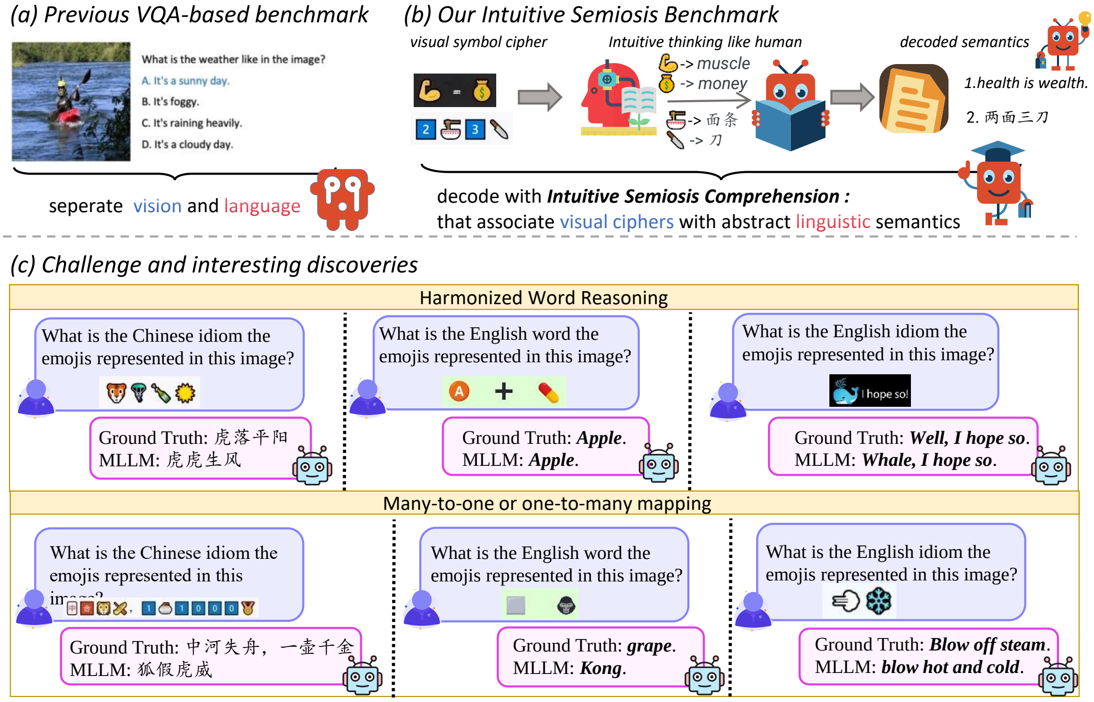

(a) Previous VQA-based benchmark
What is the weather like in the image?
A. It's a sunny day.
B. It's foggy.
C. It's raining heavily.
D. It's a cloudy day.
seperate vision and language
(b) Our Intuitive Semiosis Benchmark
visual symbol cipher
2 3
Intuitive thinking like human
-> muscle
-> money
-> 面条
-> 刀
decoded semantics
1.health is wealth.
2. 两面三刀
decode with Intuitive Semiosis Comprehension :
that associate visual ciphers with abstract linguistic semantics
(c) Challenge and interesting discoveries
Harmonized Word Reasoning
What is the Chinese idiom the emojis represented in this image?
Ground Truth: 虎落平阳
MLLM: 虎虎生风
What is the English word the emojis represented in this image?
Ground Truth: Apple.
MLLM: Apple.
What is the English idiom the emojis represented in this image?
I hope so!
Ground Truth: Well, I hope so.
MLLM: Whale, I hope so.
Many-to-one or one-to-many mapping
What is the Chinese idiom the emojis represented in this image?
Ground Truth: 中河失舟，一壶千金
MLLM: 狐假虎威
What is the English word the emojis represented in this image?
Ground Truth: grape.
MLLM: Kong.
What is the English idiom the emojis represented in this image?
Ground Truth: Blow off steam.
MLLM: blow hot and cold.

Figure 1: This figure illustrates the advantages and challenges of eWe-bench, with examples of ground truth, and MLLMs response. VQA-based benchmark example is from Li et al. (2023a).

bols into linguistic semantics while minimizing language prior interference. As shown in Figure 1, this task demands three-layer reasoning: (1) visual pattern recognition of symbolic elements, (2) combinatorial creativity to reason homophonic mappings and many-to-one correspondences, and (3) cultural interpretation of contextual relationships.

To realize this vision, we introduce **eWe-bench**, a multilingual benchmark of emoji-idiom pairs spanning Chinese and English cultural contexts. Idioms are selected from historical allusions, mythology, and conventional expressions to ensure broad linguistic and cultural coverage. We propose a data engineering framework comprising (1) retrieval of real-world emoji-text pairs, (2) text-to-emoji generation to mitigate bias, and (3) rigorous machine filtering with human validation to ensure visual sensitivity and ethical compliance.

We further design a fine-grained evaluation strategy combining automatic and human assessments. Results reveal that state-of-the-art MLLMs achieve only 3.3% compared to 67% human performance, exposing critical deficiencies in intuitive semiosis reasoning, particularly in homophonic relationships and multi-to-one mappings. A case study identifies key failures and offers insights for future improvements. By aligning MLLM evaluation with the neurocognitive of visual semiosis, we open new pathways toward human-like multimodal intelligence. Our contributions are as follows:

1. We propose a novel evaluation paradigm assessing ***visual intuitive semiosis*** of cryptographic decoding, thereby evaluating the human-like, unified visual-language comprehension of MLLMs.

2. We design a data engineering framework and construct a high-quality benchmark, eWe-bench, which minimizes the modality gap and mitigates the over-reliance on language priors.

3. Empirical evaluation reveals MLLMs' limitations in intuitive semiosis, with actionable insights applicable to cryptographical analysis and human-like intelligence.

2 Related Work

Language Model Based Cryptic Understanding Emoji can be represented by UTF-8 (Abel, 2019), and many treat emoji as text and encode them as vectors (Eisner et al., 2016). Leveraging the emoji Unicode library, numerous studies have explored emoji-text translation, including translation text into emoji (Monti et al., 2016; Leonardi, 2022; Klein et al., 2024), and bidirectional translation(Danesi, 2022). Beyond this, emoji-based sentiment analysis has become a significant area

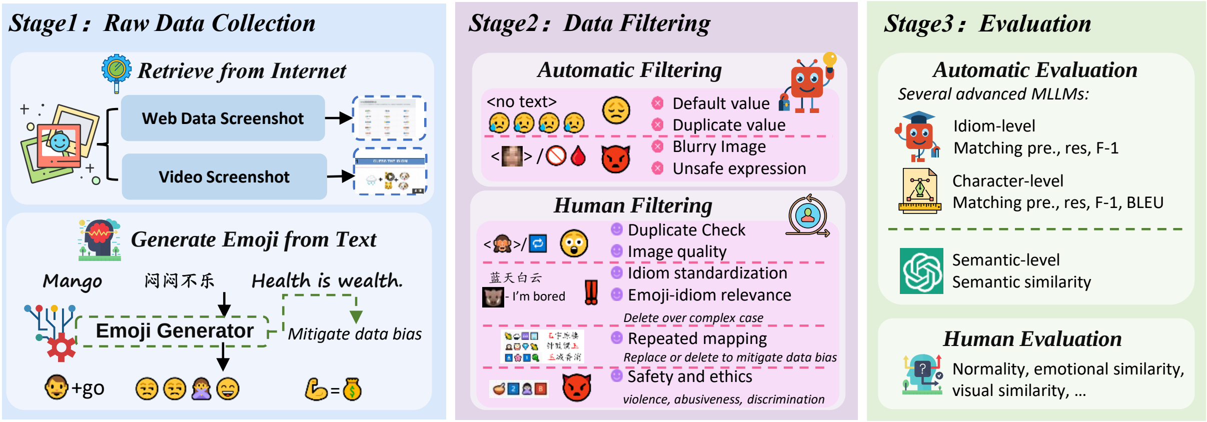


Figure 2: This figure illustrates the pipeline of eWe-bench, which is divided into raw data collection, data filtering, and evaluation.

of emoji research (Gibson et al., 2018; Chen et al., 2019, 2018; Liu et al., 2021). However, to the best of our knowledge, our eWe-bench is the first to apply the visual intuitive semiosis of emojis.

MLLMs Benchmark Earlier unified MLLMs benchmarks collect a substantial number of images and generate corresponding QA pairs to evaluate MLLMs (Fu et al., 2023a; Li et al., 2024c; Yu et al., 2024), with a focus on uniformity and objectivity, as seen in SEEDBENCH (Li et al., 2024b) and SEEDBENCH-2 (Li et al., 2023a). Recent benchmarks have started to assess different capabilities from different dimensions, including visual comprehension (Li et al., 2022b; Fu et al., 2023b; Li et al., 2024a; Tong et al., 2024; Cai et al., 2023), reasoning ability (Li et al., 2024f; Zhang et al., 2024c; Roberts et al., 2023; Huang et al., 2024), in-context learning capability (Shukor et al., 2023; Liu et al., 2023; Li et al., 2024d), hallucination challenge (Cui et al., 2023; Liu et al., 2023; Xu et al., 2025; Li et al., 2025b), and multiple domains (math, physics, music, medical, etc.)(Li et al., 2022a; Ma et al., 2022; Ye et al., 2023; Huang et al., 2023; Zhang et al., 2023; Lu et al., 2024b; Li et al., 2023c,d; Yue et al., 2024; Zhang et al., 2024b; Ye et al., 2024; Li et al., 2025c). However, most benchmarks are based on the VQA annotations and natural scenario image, rather than directly associating an abstract image with its linguistics.

3 The eWe-bench Benchmark

3.1 Task Definition

Our eWe-bench propose a benchmark that images correspond to texts with specific formats and semantics, representing a Chinese idiom word, an English word, or an English idiom sentence. Given an image of a sequence of emoji $I_i^{\text{emoji}} = \{\text{emoji}_1, \text{emoji}_2, \cdots, \text{emoji}_n\}$, eWe-bench task aims to translate emojis in images to corresponding idiom text by model F:

$$\text{Text} = F(I_i^{\text{emoji}}), \tag{1}$$

It requires not only understanding the direct corresponding text of a single emoji, but also inferring complex linguistic meaning based on intuitive semiosis capability, with detailed challenges provided in Section 3.3.

3.2 Benchmark Construction

Our data construction process involves retrieving emoji-idiom pairs from the Internet, ensuring diverse cultural representation and real-world usage. To mitigate bias, we supplement underrepresented idioms using an idiom-to-emoji generation method. A rigorous filtering pipeline, combining automatic and human review, guarantees data quality and reliability. All online data are used with proper authorization.

Raw Data Collection As shown in the Figure 2, we collect raw data through two automatic generation methods: *Retrieve from the Internet.* There are a large number of expressions of idioms based on emojis on the internet, so we retrieve the relevant web pages such as lovelyemoji [2] to get the original emoji images and the corresponding answers. *Generate Emoji Based on Text.* The quality of internet retrieval is not very high, due to 1) repetitive emoji-text pairs and recurring emoji-character mapping and 2) a relatively low number of English idioms. We utilize an existing emoji translator capable of bidirectional translation between text and

[2] https://www.lovelyemoji.com/emojicaichengyu/

Table 1: The statistics and examples of eWe-bench, where img and txt are short names of image and text.

Task	Raw data			Auto-filter	Human-filter	Example	
	Img-txt	Txt	Img	Img-txt	Img-txt	Img	Txt
Four-character idiom	2,338	2,379	2,362	2,252	1,876	😟😟❌😀	闷闷不乐
Multi-character idiom	622	641	629	576	334	❌❓3️⃣7️⃣2️⃣1️⃣0️⃣1️⃣	不问三七二十一
English Word	1,261	1,289	1,263	1,076	842	⭐🐟	starfish
English Idiom	1,237	1,254	1,244	1,182	783	💪=💰	Health is wealth.

emojis, such as emojiall[3] and additionally generate the corresponding emoji sequences, which greatly mitigates the data bias.

Automatic Filtering We employ machines and LLMs to ensure data quality. First, default values are deleted by removing incomplete emoji-idiom pairs, ensuring each emoji corresponds to a unique idiom. Duplicate pairs with identical emoji sequences are removed while retaining those with distinct emoji representations to enhance diversity. Image quality is assessed using GPT-4o to eliminate blur. Additionally, ethics checks are performed by GPT-4o, which removes emoji-idiom pairs involving violence, discrimination, and abuse. More details about automatic filtering can be found in the Appendix. A.1.

Human Filtering Human experts further refine the data that complement the automatic filtering. They assess emoji-text relevance and delete the ones that have low relevance, eliminating those with overly complex reasoning tasks. Idiom standardization is conducted to ensure alignment with human usage habits, including proper linguistic format, semantic meaning, and cultural significance. To mitigate data bias, repetitive harmonic word mappings are either replaced with alternatives or removed. Finally, safety and ethics checks are conducted to confirm the absence of inappropriate content, such as violence, sexism, stereotypes or discrimination, ensuring the dataset aligns with ethical standards. Additional information and detailed guidelines to eliminate subjectivity for human experts are provided in the Appendix A.2.

3.3 Data Statistics and Challenges

Statistics We give the statistics of eWe-bench in Table 1. After filtering, we collect high-quality data of 1,876, 334, 842, 783 emoji-texts of Chinese four-character idioms, multi-character idioms, English word and English idiom respectively (Additional statistics like word cloud in Appendix B).

Linguistic phenomena and challenges In our eWe-bench, we observe several interesting linguistic phenomena, which raise great challenges and encourage the exploration of the unified vision-language, with additional details in Appendix C.

Word Split. In English word, it is common for multiple emoji to represent one word. The word "Panda" can be split into "Pan-" and "-da," where "Pan-" corresponds to 🍳. Beyond understanding the meaning of individual emojis, the MLLMs must also remove unnecessary letters and combine the parts to infer a completely new word, raising challenge of multi-to-one mapping reasoning.

Harmonic Characters. Since it is sometimes difficult to find directly related emoji to represent, harmonic characters with similar pronunciations are often chosen to replace them. For example, "To be loaded", "To" harmonizes with "Two" 2️⃣, and "be" harmonizes with "bee" 🐝. In the Chinese idiom "难舍难离", "舍" (pronounced as "she") harmonizes with "蛇 (snake)" (pronounced as "she") of emoji 🐍, "离" harmonizes with "梨 (pear)" (pronounced as "li") of emoji 🍐.

Abstract visual Cipher Understanding. In addition to referring to the direct meanings of the emoji, it is often necessary to deeply infer the semantics of the emoji. In 🗑️❤️🍃💪"同心叶力（pull together with the same goal）", 💪 is an arm, but it does not mean "arm" in idioms. Instead, it is a very strong arm, which corresponds to "力 (power)".

Discussions of Emoji-Character Ambiguity and Ground Truth Validity Mapping emojis to homophonic characters increases task complexity and raises ground truth validity concerns. Our analysis reveals the mapping: (1) extracting the emoji's inherent meaning (e.g., 🐳 鲸 (whale)), (2) identifying potential homophones 精 (excellent)，惊 (surprise), and (3) determining the correct character based on contextual and idiomatic cues (e.g., distinguishing 惊 in "大 🐳 失色" (greatly surprised) from 精 in "🐳 才绝艳"(exceptionally talented)). While an emoji may correspond to multiple charac-

[3] https://www.emojiall.com/

Table 2: Evaluation results on Chinese idiom task. The `Word`, `Chr-2` and `Chr-1` denote the accuracy of guessing the whole word, two or more words, and one or more words correctly.

	Idiom with Four words						Idiom with Multi-words					
	Word-level			Character-level			Word-level			Character-level		
Model	**Word**	**Chr-2**	**Chr-1**	**Pre.**	**Rec.**	**F-1**	**Word**	**Chr-2**	**Chr-1**	**Pre.**	**Rec.**	**F-1**
LLaVa-1.5-7B	0.6	3.8	32.2	10.5	10.5	10.5	2.8	7.9	29.9	9.0	17.3	11.8
CogAgent-18B	0.6	4.4	34.7	11.6	11.6	11.6	3.6	8.2	30.4	8.7	14.5	10.9
Deepseek-VL-7B	0.4	2.3	25.6	6.4	6.4	6.4	1.1	4.9	29.3	8.7	10	9.3
Qwen-VL-7B	0.5	4.7	30.2	8.7	8.7	8.7	2.4	9.8	31.7	9.1	16.9	11.8
Qwen2.5-VL-7B	1.8	8.6	**40.7**	**12.3**	**12.3**	**12.3**	5.6	11.2	**34.8**	**11.3**	**19.2**	**14.2**
InternVL-2-8B	0.8	6.3	37.8	9.1	9.1	9.1	3.4	8.3	29.4	8.9	15.6	11.3
Claude-3.5	1.3	6.7	23.3	8.0	8.0	8.0	1.4	2.9	7.1	6.0	9.7	7.4
GPT-4v	0.7	1.3	22.1	5.8	5.8	5.8	1.1	6.8	28.4	3.7	9.1	5.3
GPT-4o	**3.3**	**8.7**	27.5	10.7	10.7	10.7	**9.1**	**13.6**	27.3	7.5	**18.1**	10.6

ters, the final mapping remains contextually well-defined. In our eWe-bench, we model this process by encoding emoji sequences with their structure, ensuring a unique and unambiguous ground truth. More details about the emoji sequence and structures ensure the validity of ground truth in Appendix C.4.

3.4 Evaluation Metrics

The researcher can gain details of metric computation and insights that the MLLMs capabilities revealed by each metric in Appendix. D. For ***automatic evaluation*** (Appendix D.2): 1) For the whole idiom, we compute the Precision, Recall, and F-1 value from the word/sentence-level that exactly matches the ground truth, which is the most direct metric to show whether MLLMs understands the abstract linguistic semantics of the image. 2) For characters in the idiom, we compute the Precision, Recall, F-1. and BLEU value that matches the ground truth from the character/word-level without considering the structural correspondence, which is aligned with fine-grained abstract linguistic semantics of image understanding. 3) In addition to the direct matching, we calculate the semantic similarity between the predicted answers and ground truth using GPT-4o. We also perform ***human evaluation*** of the model outputs (Appendix D.3) and discussion of future metrics (Appendix D.4).

4 Experiment Results

4.1 Baselines and Implementation

We select commercial Claude-3.5-sonnet-20241022, gpt-4-vision-preview and GPT-4o-20240513 to evaluate the eWe-bench benchmark. For a richer evaluation, we select a series of open-source MLLMs for testing. These include: 1) Qwen-VL-7B and Qwen2.5-VL-7B (Bai et al., 2023), DeepSeek-VL-7B (Lu et al., 2024a), which have good **Chinese language** support; 2) LLaVa-1.5-7B (Li et al., 2023b), CogAgent-18B (Hong et al., 2023), InternVL-2-8B which have good **visual comprehension** capabilities. We provide details of the baselines, implementation details, and the prompt in Appendix E.

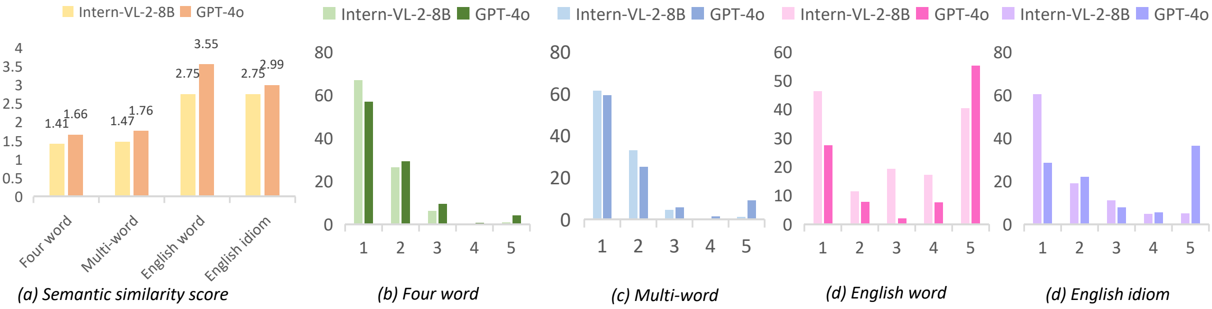
4.2 Automatic Evaluation Results

Emoji to Chinese Idiom We evaluate four-character and multi-character idioms shown in Table 2 (error bar in Appendix F.1). We observe that all the MLLMs perform poorly: The latest model, GPT-4o, achieves accuracy scores of 3.3 and 5.0 at the word level for both tasks. However, even the strongest open source model Qwen2.5-VL still has a big gap with it, which proves that the different strengths of the models can still be observed on our more difficult benchmark. The accuracy at the Chr-1 is significantly higher, indicating that MLLMs are equipped with the basic translations of text corresponding to individual emojis, but have limited visual intuitive semiosis capability to further infer the corresponding linguistic meanings based on the relevant emoji context, especially for the harmonization reasoning with detailed analysis in Sec. 4.5. Thus, our eWe-bench is challenging for MLLMs to decode the visual ciphers.

Emoji to English Word and English Idiom MLLMs's performance is higher compared to the two Chinese tasks ((error bar in Appendix F.1)). In Table 3, GPT-4o achieves impressive F-1 of 55.8

Table 3: Evaluation on English word and idiom task. B-1 and B-2 denote the BLEU-1 and BLEU-2 respectively.

	English Word						English Idiom							
	Word-level			Character-level			Sentence-level			Word-level				
Model	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**	**B-1**	**B-2**
LLaVA-1.5-7B	30.1	30.1	30.1	54.6	55.7	55.1	14.4	14.4	14.4	19.7	21.3	20.5	19.7	16.4
CogAgent-18B	29.8	29.8	29.8	52.8	51.9	52.3	13.2	13.2	13.2	18.3	19.5	18.9	18.3	15.2
Deepseek-VL-7B	23.2	26.3	24.7	46.2	47.5	46.8	11.9	11.9	11.9	15.1	14.6	14.8	15.1	11
Qwen-VL-7B	28.6	29.1	28.8	51.2	50.4	50.8	12.1	12.1	12.1	17.1	12.5	14.4	17.1	11.3
Qwen2.5-VL-7B	33.7	33.7	33.7	57.6	58.9	58.2	18.6	18.6	18.6	23.5	21.6	22.5	20.7	19.2
InternVL-2-8B	31.1	31.1	31.1	56.6	57.2	56.9	15.3	15.3	15.3	19.3	22.1	20.6	18.4	16.1
Claude-3.5	42.3	42.3	42.3	63.9	73.8	68.5	29.8	29.8	29.8	48.0	42.7	45.2	42.3	39.7
GPT-4v	38.5	38.5	38.5	60.3	69.2	64.4	26.4	26.4	26.4	41.1	43.1	42.1	39.4	37.5
GPT-4o	**55.8**	**55.8**	**55.8**	**68.5**	**77.5**	**72.7**	**35.2**	**35.2**	**35.2**	**46.8**	**47.3**	**47.0**	**45.0**	**41.6**



(a) Semantic similarity score

(b) Four word

(c) Multi-word

(d) English word

(d) English idiom

Figure 3: The results of semantic similarity scores and distribution of Intern-VL-2 and GPT-4o.

and 35.2 at the word and sentence levels, in English word and idiom respectively. This is likely because the model has encountered more similar English texts during training, making it more adept at reasoning about English words. However, MLLMs always suffer from hallucination problems when decoding visual cipher emojis, with detailed analysis in Sec. 4.5. When they catch a linguistic meaning of a single emoji, they quickly focus on the word or idiom related to this emoji from the inner knowledge they have, and ignore the relevant context of the emojis. Based on our eWe-bench, the community can explore the hallucination problem and improve the inference ability.

Evaluation of the semantic similarity We further compute the semantic similarity between the responses and the ground truth, applying LLM to score from 1 to 5. As shown in Figure 3, the average scores are low, with the English task significantly higher than those on the Chinese task. When carefully observing the distribution, we observe that 1)for the Chinese task, most of the scores are concentrated in 1 and 2, indicating the poor performance of MLLMs; 2)while for the English task, most of the scores are concentrated in 1 and 5, demonstrating that the MLLMs can either predict the answer correctly, or get irrelevant answers.

4.3 Further Exploration

Exploration with In-context Learning In addition to evaluating the direct inference abilities of MLLMs, we further explore their performance using in-context learning. We select the open-source Qwen-VL and the closed-source GPT-4o, evaluating each task with 3, 5, and 7 context examples, as shown in Table 4 and Table 5. MLLMs improve across various tasks with the addition of contextual examples, indicating the high quality of our Emoji2Idiom that the randomly chosen examples can improve the performance a lot. However, in the Chinese task, performance decreases when using too many samples (7 in-context examples). This decline indicates that MLLMs learn incorrect mappings in this complex task and suffer from hallucination issues.

Exploration with Chain-of-Thought We investigate the enhancement of CoT, prompting the MLLMs to think the fundamental meaning and reason homophones like humans, with the details in Appendix E.4.2. In Figure 4, by mimicking human reasoning, the CoT design enables the MLLMs to

Table 4: Exploration on in-context learning in Chinese idiom tasks.

	Idiom with Four words						**Idiom with Multi-words**					
	Word-level			**Character-level**			**Word-level**			**Character-level**		
Model	**Word**	**Chr-2**	**Chr-1**	**Pre.**	**Rec.**	**F-1**	**Word**	**Chr-2**	**Chr-1**	**Pre.**	**Rec.**	**F-1**
Qwen-VL-7B	0.5	4.7	30.2	8.7	8.7	8.7	2.4	9.8	31.7	9.1	16.9	11.8
+3 in-context example	0.5	5.1	31.3	9.3	9.3	9.3	2.2	10.1	28.6	10.4	13.1	11.6
+5 in-context example	0.6	5.3	31.6	9.4	9.4	9.4	3.3	12.3	32.1	11.7	16.9	13.8
+7 in-context example	0.5	4.9	32.1	9.4	9.4	9.4	2.8	10.7	31.4	11.4	15.4	13.1
GPT-4o	3.3	8.7	27.5	10.7	10.7	10.7	9.1	13.6	27.3	7.5	18.1	10.6
+3 in-context example	2.6	11.3	33.9	12.6	12.6	12.6	9.5	23.8	36.9	17.0	21.9	19.1
+5 in-context example	3.5	12.2	35.7	13.7	13.7	13.7	13.1	27.4	42.9	20.7	29.1	24.2
+7 in-context example	3.5	8.7	31.3	12.0	12.0	12.0	10.7	19.0	34.5	16.2	23.1	19.0

Table 5: Exploration on in-context learning in English tasks.

	English Words						**English Idiom**					
	Word-level			**Character-level**			**Sentence-level**			**Word-level**		
Model	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**	**Pre.**	**Rec.**	**F-1**
Qwen-VL	28.6	29.1	28.8	51.2	50.4	50.8	12.1	12.1	12.1	17.1	12.5	14.4
+3 in-context example	29.3	29.3	29.3	53.6	52.1	52.8	12.0	12.0	12.0	17.6	17.9	17.7
+5 in-context example	30.6	30.6	30.6	55.9	54.2	55.0	13.0	13.0	13.0	18.9	18.4	18.6
+7 in-context example	32.5	32.5	32.5	57.8	55.7	56.7	15.2	15.2	15.2	19.7	20.2	19.9
GPT-4o	55.8	55.8	55.8	68.5	77.5	72.7	35.2	35.2	35.2	46.8	47.3	47.0
+3 in-context example	57.6	57.6	57.6	72.3	75.0	73.6	36.3	36.3	36.3	47.6	47.3	47.4
+5 in-context example	54.5	54.5	54.5	77.5	79.0	78.2	37.4	37.4	37.4	48.2	50.0	49.1
+7 in-context example	60.6	60.6	60.6	79.4	73.9	76.5	38.5	38.5	38.5	49.5	50.5	50.0

produce better answers without additional training, improving accuracy at both character and word levels while improving semantic and visual similarity. This demonstrates the method's effectiveness and the high quality of our data. We provide some further experimental results and insights of this harmonic symbol chain od thought reasoning in Appendix. H.1.

Effect of Input Length and Emoji Representation We explore the impact of **emoji sequence length**. The balanced lengths of emoji sequence of Chinese four-character idioms and English words lead to minimal impact from resizing images. However, Chinese multi-character English idioms have longer sequences (averaging 7.48 and 5.32), tending to suffer from distortion due to resizing and performance degradation. Additionally, we investigate **emoji representation differences across platforms** (Windows, Mac, Android) and confirm that our unified rendering strategy on Windows does not significantly affect evaluation results. For a detailed discussion and experimental findings, please refer to the Appendix F.3.

4.4 Human Evaluation

Human Performance We invite human experts to participate and assess the task's difficulty, thereby determining the upper limit of machine performance on this benchmark. The same evaluation metrics tests humans, and task complexity is rated on a scale from one (very easy) to five (very difficult), with more details on the evaluation and detailed results are provided in Appendix D.3 and F.4. Figure 5 show that MLLMs still has significant room for improvement, and our eWe-bench presents significant challenges.

Human Evaluation on MLLMs We conduct a human evaluation, assessing normality, semantic and emotional similarity to the ground truth, visual resemblance to emojis, and text fluency (details in Appendix F.5). As shown in Figure 6, Chinese performance falls behind English, aligning with automatic evaluation. While all MLLMs generate standardized idioms well, lower semantic and visual similarity highlight challenges in semiosis reasoning. Notably, GPT models excel in visual un-

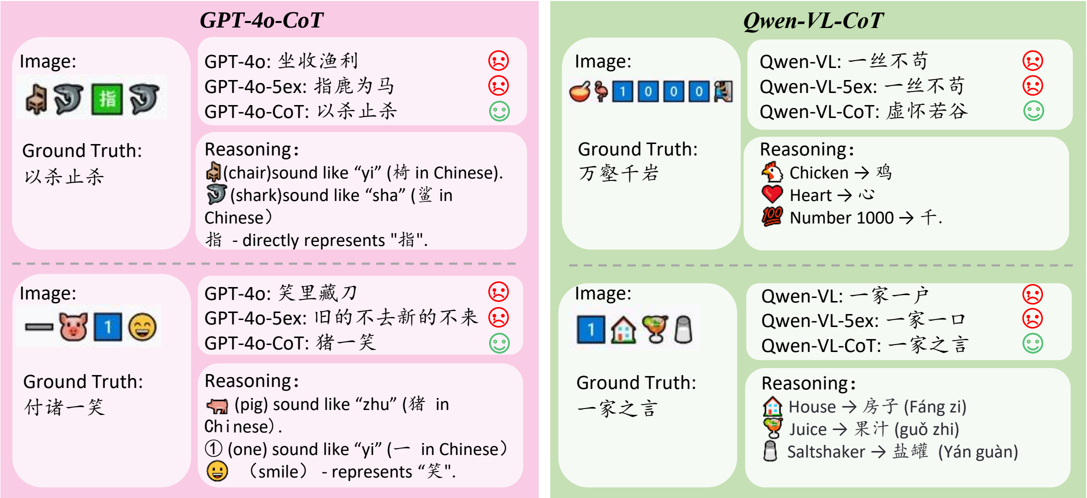


Figure 4: The results of the base, ICL, and CoT approach, with the reasoning process of MLLMs.

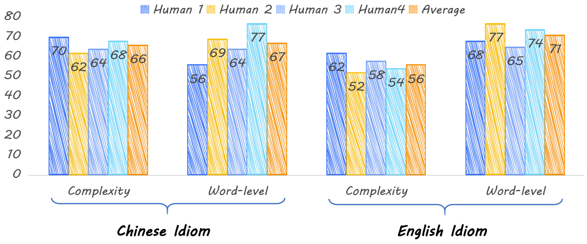
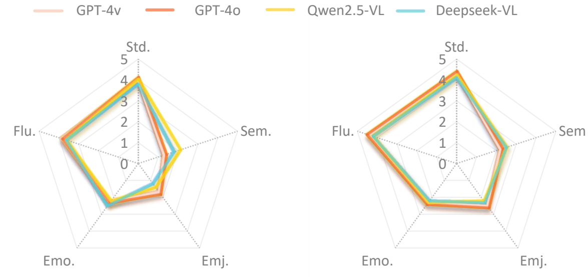


Figure 5: Human performance on Chinese idiom task, where we map the score to the 1-100 interval.



(a) Human evaluation on Chinese Idiom

(b) Human evaluation on English Idiom

Figure 6: Human evaluation. The std., sem, emo., vis., and flu. denote the standardization, semantic and emotional similarity to ground truth, visual similarity to the image, and fluency.

derstanding, while open-source models, especially Qwen2.5-VL, perform better in text semantics, offering insights for future research.

4.5 Case Study

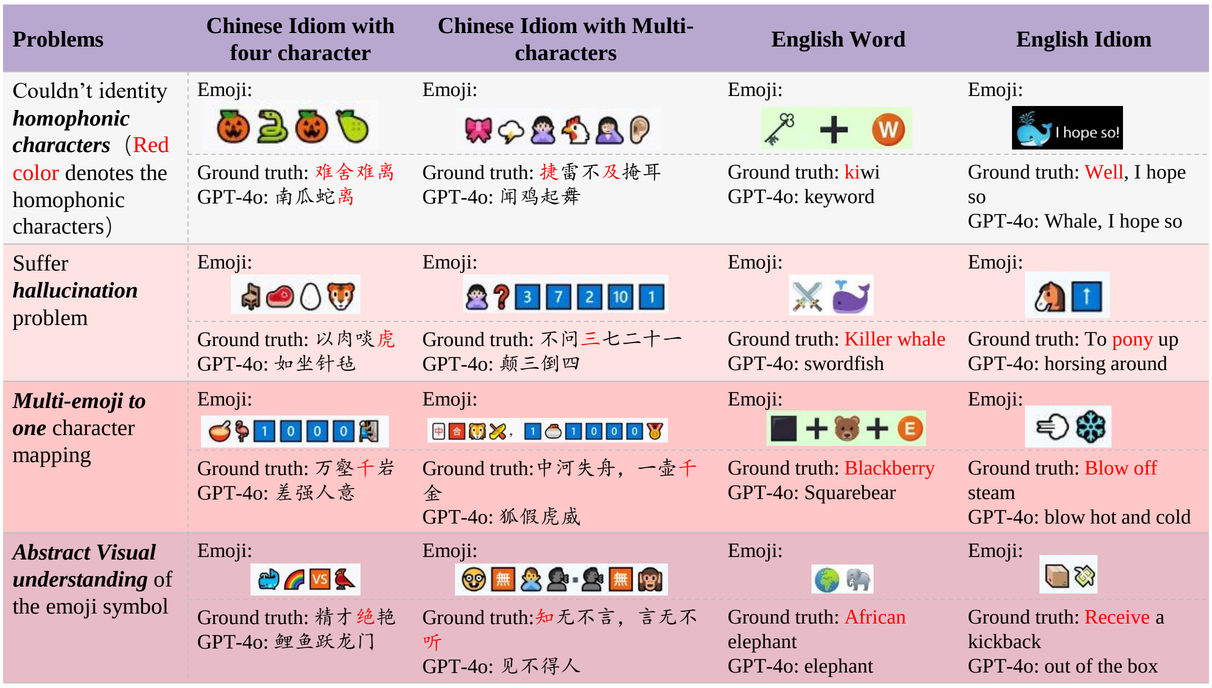
We present a case study highlighting challenges (more examples in Appendix G) and insights for future intuitive semiosis MLLMs in Appendix H.2.

Harmonization Problem MLLMs often fail on harmonization problems. As shown in Figure 7, 🐍 - "蛇 (snake, sound like "she")" homonym to "舍 (leave, sound like "she")" but the MLLMs fails to recognize. In the English idiom, 🐋 - "whale" homonym to "well". Our eWe-bench includes many harmonic character phenomena. Current MLLMs are not yet capable of effectively capturing visual context and reasoning with harmonic words, and struggle with our challenging eWe-bench.

Hallucination Problem In Figure 7, the model recognizes 🐎 and immediately outputs "horsing around", without considering the surrounding emojis. Another example shows 3️⃣ directly associated with the idiom "颠三倒四 (disorderly)", both containing the number 3. That is due to the hallucination problem. MLLMs often think narrowly, focusing only on words or idioms directly related to a single emoji. Our eWe-bench looks forward to further exploration of the poor performance of MLLMs that we have discovered.

Multi-emoji to One Character Mapping. eWe-bench presents a huge challenge on this mapping issue, where MLLMs fail to perform a multi-to-one or one-to-multi mapping. For example, 1️⃣ 0️⃣ 0️⃣ 0️⃣ are four emojis, but the model does not successfully combine them into one character "千 (one thousand)", which greatly inspire the application of cryptography.

Abstract Visual Understanding MLLMs struggle to align emoji semantics with intricate meanings when it comes to deep comprehension. For example, in "receive a kickback", the model simply captures 📦, the meaning of "box", and interprets it as "out of the box", but does not combine the



Problems	Chinese Idiom with four character	Chinese Idiom with Multi-characters	English Word	English Idiom
Couldn't identity ***homophonic characters***（Red color denotes the homophonic characters）	Emoji: Ground truth: 难舍难离 GPT-4o: 南瓜蛇离	Emoji: Ground truth: 捷雷不及掩耳 GPT-4o: 闻鸡起舞	Emoji: Ground truth: kiwi GPT-4o: keyword	Emoji: I hope so! Ground truth: Well, I hope so GPT-4o: Whale, I hope so
Suffer ***hallucination*** problem	Emoji: Ground truth: 以肉啖虎 GPT-4o: 如坐针毡	Emoji: Ground truth: 不问三七二十一 GPT-4o: 颠三倒四	Emoji: Ground truth: Killer whale GPT-4o: swordfish	Emoji: Ground truth: To pony up GPT-4o: horsing around
Multi-emoji to one character mapping	Emoji: Ground truth: 万壑千岩 GPT-4o: 差强人意	Emoji: Ground truth:中河失舟，一壶千金 GPT-4o: 狐假虎威	Emoji: Ground truth: Blackberry GPT-4o: Squarebear	Emoji: Ground truth: Blow off steam GPT-4o: blow hot and cold
Abstract Visual understanding of the emoji symbol	Emoji: Ground truth: 精才绝艳 GPT-4o: 鲤鱼跃龙门	Emoji: Ground truth:知无不言，言无不听 GPT-4o: 见不得人	Emoji: Ground truth: African elephant GPT-4o: elephant	Emoji: Ground truth: Receive a kickback GPT-4o: out of the box

Figure 7: Four typical challenges the GPT-4o suffer in our eWe-bench.

attributes of "receiving something" with the hint of money to generate the correct answer. Our eWe-bench highly focuses on this deeper understanding, exploring the capabilities of MLLMs.

5 Conclusion

We propose the eWe-bench benchmark, containing emoji visual ciphers decoding tasks including Chinese idioms, English words, and English idioms, which provides a novel way to measure the visual intuitive semiosis ability of MLLMs to directly associate the image with its abstract linguistic semantics. We design a data engineering framework that performs data collection, data filtering, and evaluation, contributing to validating the unified high-level vision-language understanding and reasoning like human intuitive cognition. We evaluate advanced open-source and closed-source MLLMs with our eWe-bench, analyze the performance, conduct several explorations, and highlight future research directions with further case study.

Limitations

In our proposed eWe-bench, we select the two most commonly used emoji-to-idiom text types: Chinese and English, highlighting the challenges of our benchmark. Additionally, more languages, such as Japanese, Korean, French, German, and Arabic, can be incorporated to further evaluate MLLMs' ability to understand the correlation between images and linguistic semantics.

Ethics Statement

We introduce a novel benchmark, eWe-bench, incorporating a thorough description of data collection, annotation, and filtration processes. We emphasize that the dataset's creation adheres strictly to ethical guidelines. Great care has been taken to uphold ethical standards in the dataset, employing anonymization, desensitization, and data cleaning. The samples pose no risk to public welfare. For all data sourced from these websites, we contact the site administrators to obtain permission for data usage. Additionally, we sign intellectual property sharing agreements with them to ensure compliance with ethical and legal standards. Hence, the innovative research directions and tasks proposed are ethically harmless to society.

Acknowledgement

This research is supported by Key-Area Research and Development Program of GuangdongProvince No.2024B1111100001.

References

Jonathan E Abel. 2019. Not everyone s: Or, the question of emoji as ‘universal’ expression. In *Emoticons, Kaomoji, and Emoji*, pages 25–43. Routledge.

Giuseppe Ateniese, Carlo Blundo, Alfredo De Santis, and Douglas R Stinson. 1996. Visual cryptography for general access structures. *Information and computation*, 129(2):86–106.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Rizhao Cai, Zirui Song, Dayan Guan, Zhenhao Chen, Xing Luo, Chenyu Yi, and Alex Kot. 2023. Benchlmm: Benchmarking cross-style visual capability of large multimodal models. *arXiv preprint arXiv:2312.02896*.

Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter sentiment analysis via bi-sense emoji embedding and attention-based lstm. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 117–125.

Zhenpeng Chen, Sheng Shen, Ziniu Hu, Xuan Lu, Qiaozhu Mei, and Xuanzhe Liu. 2019. Emoji-powered representation learning for cross-lingual sentiment classification. In *The world wide web conference*, pages 251–262.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Marcel Danesi. 2022. Emotional wellbeing and the semiotic translation of emojis. In *Exploring the Translatability of Emotions: Cross-Cultural and Transdisciplinary Encounters*, pages 323–344. Springer.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Changde Du, Kaicheng Fu, Jinpeng Li, and Huiguang He. 2023. Decoding visual neural representations by multimodal learning of brain-visual-linguistic features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10760–10777.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.

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

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023b. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

Will Gibson, Pingping Huang, and Qianyun Yu. 2018. Emoji and communicative action: The semiotics, sequence and gestural actions of ‘face covering hand’. *Discourse, Context & Media*, 26:91–99.

J Allan Hobson and Edward F Pace-Schott. 2002. The cognitive neuroscience of sleep: neuronal systems, consciousness and learning. *Nature Reviews Neuroscience*, 3(9):679–693.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for GUI agents. *CoRR*, abs/2312.08914.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11514–11525. Association for Computational Linguistics.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. 2024. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10186–10197. ELRA and ICCL.

Muhammad Nazrul Islam and Harry Bouwman. 2016. Towards user–intuitive web interface sign design and evaluation: A semiotic framework. *International Journal of Human-Computer Studies*, 86:121–137.

Lars Henning Klein, Roland Aydin, and Robert West. 2024. Emojinize: Enriching any text with emoji translations. *arXiv preprint arXiv:2403.03857*.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Jiayi Kuang, Jingyou Xie, Haohao Luo, Ronghao Li, Zhe Xu, Xianfeng Cheng, Yinghui Li, Xika Lin, and Ying Shen. 2024. Natural language understanding and inference with mllm in visual question answering: A survey. *arXiv preprint arXiv:2411.17558*.

Vanessa Leonardi. 2022. Communication challenges and transformations in the digital era: emoji language and emoji translation. *Language and Semiotic Studies*, 8(3):22–44.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal llms with generative comprehension. In *CVPR*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025a. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yangning Li, Tingwei Lu, Hai-Tao Zheng, Yinghui Li, Shulin Huang, Tianyu Yu, Jun Yuan, and Rui Zhang. 2024c. MESED: A multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8697–8706. AAAI Press.

Yinghui Li, Haojing Huang, Jiayi Kuang, Yangning Li, Shu-Yu Guo, Chao Qu, Xiaoyu Tan, Hai-Tao Zheng, Ying Shen, and Philip S. Yu. 2025b. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023c. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Jiayi Kuang, Haojing Huang, Zhikun Xu, Xinnian Liang, Yi Yu, Wenlian Lu, Yangning Li, Xiaoyu Tan, Chao Qu, Ying Shen, Hai-Tao Zheng, and Philip S. Yu. 2025c. One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical llms. *CoRR*, abs/2502.10454.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1077–1086. ACM.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Shang Qin, Jingheng Ye, Shirong Ma, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S. Yu. 2024d. Rethinking the roles of large language models in chinese grammatical error correction. *CoRR*, abs/2402.11420.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Shirong Ma, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2024e. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8656–8668. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024f. When llms meet cunning texts: A fallacy understanding benchmark for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023d. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024a. 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. 2024b. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. In *ICLR*.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7978–7993. Association for Computational Linguistics.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Yangning Li, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 576–589. Association for Computational Linguistics.

Johanna Monti, Federico Sangati, Francesca Chiusaroli, Benjamin Martin, Mansour Sina, et al. 2016. Emojitalianobot and emojiworldbot-new online tools and digital environments for translation into emoji. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016)*.

Pierre Moulin and Joseph A O'Sullivan. 2003. Information-theoretic analysis of information hiding. *IEEE Transactions on information theory*, 49(3):563–593.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Jonathan Roberts, Timo Lüddecke, Rehan Sheikh, Kai Han, and Samuel Albanie. 2023. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. *arXiv preprint arXiv:2311.14656*.

Luca M Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. 2025. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, pages 1–11.

Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2023. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *arXiv preprint arXiv:2310.00647*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

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

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 2024. Dynamic demonstration retrieval and cognitive understanding for emotional support conversation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 774 – 784, New York, NY, USA. Association for Computing Machinery.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! A chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10435–10448. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024. EXCGEC: A benchmark of edit-wise explainable chinese grammatical error correction. *CoRR*, abs/2407.00924.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19458–19467. AAAI Press.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Ding Zhang, Yinghui Li, Qingyu Zhou, Shirong Ma, Yangning Li, Yunbo Cao, and Hai-Tao Zheng. 2023. Contextual similarity is more valuable than character similarity: An empirical study for chinese spell checking. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, et al. 2024b. Cm-mmu: A chinese massive multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2401.11944*.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024c. Benchmarking large multimodal models against common corruptions. In *NAACL*.

A Additional Details of Data Filtering

A.1 Automatic Filtering

In this phase, we mainly utilize machines and large language models to filter large-scale data, which includes the following steps in total:

1. Deletion of default values. We utilize a machine to automatically remove incomplete emoji-idiom pairs, including those with missing corresponding emoji images and those with missing corresponding idioms. It is guaranteed that each emoji image corresponds to the standard idiom answer one by one.

2. Removing Duplicate Values. We utilize the machine to automatically remove duplicate emoji-idiom pairs. Here, we only need to remove the emoji-idiom pairs corresponding to identical emoji sequences while retaining the pairs with the same idiom text but corresponding to different emoji representations, which helps to enhance the diversity of the dataset. Note that we will first filter the pairs corresponding to the same idiom text by the machine with additional labels, and make a manual decision on whether to perform the deletion in the next stage of manual filtering.

3. Image Quality Check. We utilize LLM (specifically GPT-4o is used), to perform image quality checking, which entails marking and removing: images that are too blurry and those that do not meet the ethical norms (images that contain elements of violence, abusive language, discrimination, etc.) along with their corresponding idioms.

4. Text Ethics Checking. We utilize LLM (specifically GPT-4o) to perform text ethics checking, which involves tagging and deleting idiom with elements of violence, discrimination, abuse, etc. For example, "红颜祸水" is a sexist idiom, and we will delete its corresponding emoji-idiom pair.

A.2 Human Filtering

In this phase, we invited human experts in Chinese and English languages to perform manual data filtering, which included the following steps in total:

1. Duplicate value checking: for the automatic filtering phase, the machine flags a portion of emoji-idiom pairs where the text is the same but the corresponding images are not the same. the human expert needs to further check whether the emoji expressions here are really different. For the pairs with identical emoji images, the human expert will delete them.

2. Image quality check: Human experts further check whether the emoji images are unclear and illegible, and remove the illegible images.

3. Idiom standardization check: Human experts need to check whether the idiom text expression is standardized, including the format of the idiom, whether it has a specific linguistic meaning, and whether it is in line with common human usage, etc., to ensure that our dataset meets the real-world usability. For example, for the idiom "blue sky and white clouds", although it is a four-word idiom that conforms to the norms of human usage, it does not have a specific allusion, mythological story, traditional story background, or special semantic meaning, and does not belong to the standard idioms. For example, although "流水高山" is a four-letter word with a specific historical background, people more often use the expression "high mountains and flowing water". Therefore, "流水高山" is not an expression that conforms to human language usage and will be deleted.

4. Emoji and Idiom Relevance Check: Since in emoji to idiom expression, many times the representation of harmonic characters will be utilized, which will increase the difficulty of emoji comprehension and the difficulty of generating the final idioms. Human experts will evaluate the relevance of emoji to idioms:

   - If too many or too complex harmonic characters are used with the emoji representation, at this time the task will be too difficult for not only MLLMs but also humans to understand. At this point, the human expert will consider the relevance of this emoji sequence to the idiom to be too low and delete the emoji-idiom pair.
   - It is noteworthy that we conducted an evaluation of human ability on this benchmark in Sec. 4.4 and found that humans achieved an average score of 66.5 on the word-level accuracy of Chinese idioms. This score demonstrates both that our dataset is challenging and that the task is accomplishable, and that there is still much room for improvement in the performance of the current MLLMs on this task.

5. Repeated harmonic word mapping check: due to the limited expression of emoji, when using emoji to replace textual expressions, harmonic words are often used to find the corresponding emoji for expression. eWe-bench also has a large number of harmonic words. However, we found that if the mapping of the same emoji corresponding to a certain harmonic word occurs too many times, it may cause data bias to LLM in subsequent training, i.e., when LLM sees this emoji it automatically thinks of this harmonic word that occurs multiple times. To mitigate the bias caused by this harmonic word mapping, we performed:

   - Count the repeated emoji-character harmonic word mappings, and when there are more than ten occurrences, we manually replace the expression of the emoji (find other harmonic word counterparts to replace the original repeated emoji), or just delete the redundant emoji-idiom pair.
   - In addition, we also considered this issue during the original data collection. Our retrieval and collection in different sources of the original emoji database can reduce this duplicate mapping. We also take different generation methods when manually constructing text-to-emoji data, which also helps to increase the diversity of harmonic word mappings.

6. Safety and Ethics Check: Based on the automatic detection, the human experts further conducted a safety and ethics check of the emoji images and idiom text, checking whether there are any issues such as violent gore, abusive language, sexism, racial discrimination, stereotyping, and so on, in the data.

To eliminate subjectivity in manual filtering, we provide annotators with detailed guidelines as shown in Figure 8 and 9, including scoring criteria for each item (1-5 points) covering idiomatic normality, graphic consistency, image legibility, repetition mapping, and ethical safety checks. We also provide at least three examples for each item. For ethical safety checks, we distinguish between subcategories such as violence, abusive language, gender discrimination, stereotyping, and racial discrimination. We provide examples at both the emoji and text levels to guide judgments. The annotators are student volunteers who are native speakers of Chinese and English.

B Additional Details of Data statistics

In addition to the numerical statistics, we further do some statistics to better show our eWe-bench.

Word Frequency and Word Cloud Statistic of Chinese idiom To better present our dataset, we perform word frequency statistics on Chinese idioms and display the word cloud and word rectangle tree graphs, as shown in Figure 10. We first perform word frequency statistics on all characters, filter out the top 1,000 characters, and discard low-frequency words. From the filtered top 1,000 characters, we conduct lexical analysis and plot word cloud and word rectangle diagrams for adjective and adverbial morphemes, noun morphemes, and verb morphemes, respectively.

Word Frequency and Word Cloud Statistic of English idiom Similarly, we perform word frequency statistics on English idioms and display word cloud maps with word rectangle tree diagrams, as shown in Figure 11. We first perform word frequency statistics on all words, filter out the top 180 words, and discard low-frequency words. From the filtered top 180 words, we create word cloud maps with word rectangle mapping.

C Additional Details of Data Attributes and Linguistic Phenomenon

C.1 Chinese Idiom Task

Harmonization Word Since it is sometimes difficult to find directly related emoji to represent, harmonic characters with similar pronunciations are often chosen to replace them. For example, Usually, for characters that can't be represented directly by emoji, we will first search for their harmonized characters, then find an emoji that can directly represent the harmonized character, and replace it with this emoji. For example, "捷" does not have a direct emoji, but it harmonizes with 结", which corresponds to "bow" 🎀, and so, we select 🎀 chosen to represent the character "捷". There are a large number of harmonic characters in our data. This poses a great challenge to MLLMs's understanding and reasoning ability. The reasoning of harmonic words needs the help of related contexts, and in our data scenario, the model is required to analyze the context of emoji in depth instead of understanding individual emoji alone. The understanding of these harmonics usually requires the model to synthesize the relevant context of the emoji, to reason out the correct expression of the harmonized words.

Abstract visual Emoji Understanding. The model shows better performance in simply recognizing the shallow meanings of individual emoji, but in Abstract visual in-depth understanding, it is difficult for the model to work with the contextual emoji information to get the real corresponding relevant emoji meanings. For example, 🆚 means match, PK, duel, competition, and so on. In Chinese, "决" represents duel, and then harmonized to "绝" to get the idiom "精才绝艳". In "African elephant", the superficial meaning of the emoji is the earth, but further combined with the specific location of the earth map in the figure and the hint of an elephant, the emoji represents the African elephant. Abstract visual understanding in conjunction with its textual meaning to further reason about the correct answer. In 🗑️❤️🍃💪"同心叶力（pull together with the same goal）", 💪 is an arm, but it does not mean "arm" in idioms. Instead, it is a very strong arm, which corresponds to "力 (power)".

Chinese Idiom Format Chinese idioms are a special kind of words, which often have specific formats and semantic information, so they cannot directly translate the meaning of a single emoji and concatenate words into sentences. The most common format is four-character idioms, which often come from ancient Chinese myths, historical stories, classics, etc., consisting of four Chinese characters, with a Chinese literary style, and often a symmetrical structure. In addition, multi-character idioms, although far fewer in number than four-character idioms, are equally important components. Some of them have less than four words (e.g., three-character idioms) and some have more than four words. Generally speaking, whether it is a four-character idiom or a multi-character id-

Guideline of Human Filtering of Data - 1

The purpose of this work is to screen out emoji-text pairs that do not comply with the rules. For each indicator, there will be a corresponding criterion and examples, which you will need to score the emoji-text pairs, and only the pairs that meet the requirements of each indicator can be retained.

Case

Image: 🎀🌩🙅🐔🤦👂 **Text：捷雷不及掩耳**	Normality: 5 Consistency: 4 legibility: 4 Emoji Ethical security: 4 Text Ethical security: 5

Metrics

➢ **Normality:** Whether the text conforms to the idiom‘s specifications. This includes whether it has historical allusions and specific cultural backgrounds, or does not conform to human usage habits

Options	1. Completely non-standard 2. Mostly non-standard 3. Fairly standard 4. Mostly standard 5. Completely standard
Examples	1. “朝三暮四” shows completely standard to the normality. 2. “蓝天白云” shows completely non-standard. 3. “红红火火” shows mostly non-standard.

➢ **Consistency**: The consistency of the emoji and the image is scored, and the higher the consistency of the example, the easier it is to get the final translation result

Options	1. Completely inconsistent 2. Mostly inconsistent 3. Fairly consistent 4. Mostly consistent 5. Completely consistent
Examples	1. 2️⃣🍜3️⃣🔪 - “两面三刀” pair is mostly consistent. 2. 🌄👂↩️👂 - “出尔反尔” pair is mostly inconsistent. 3. 🐗 - “I’m bored.” pair is completely inconsistent.

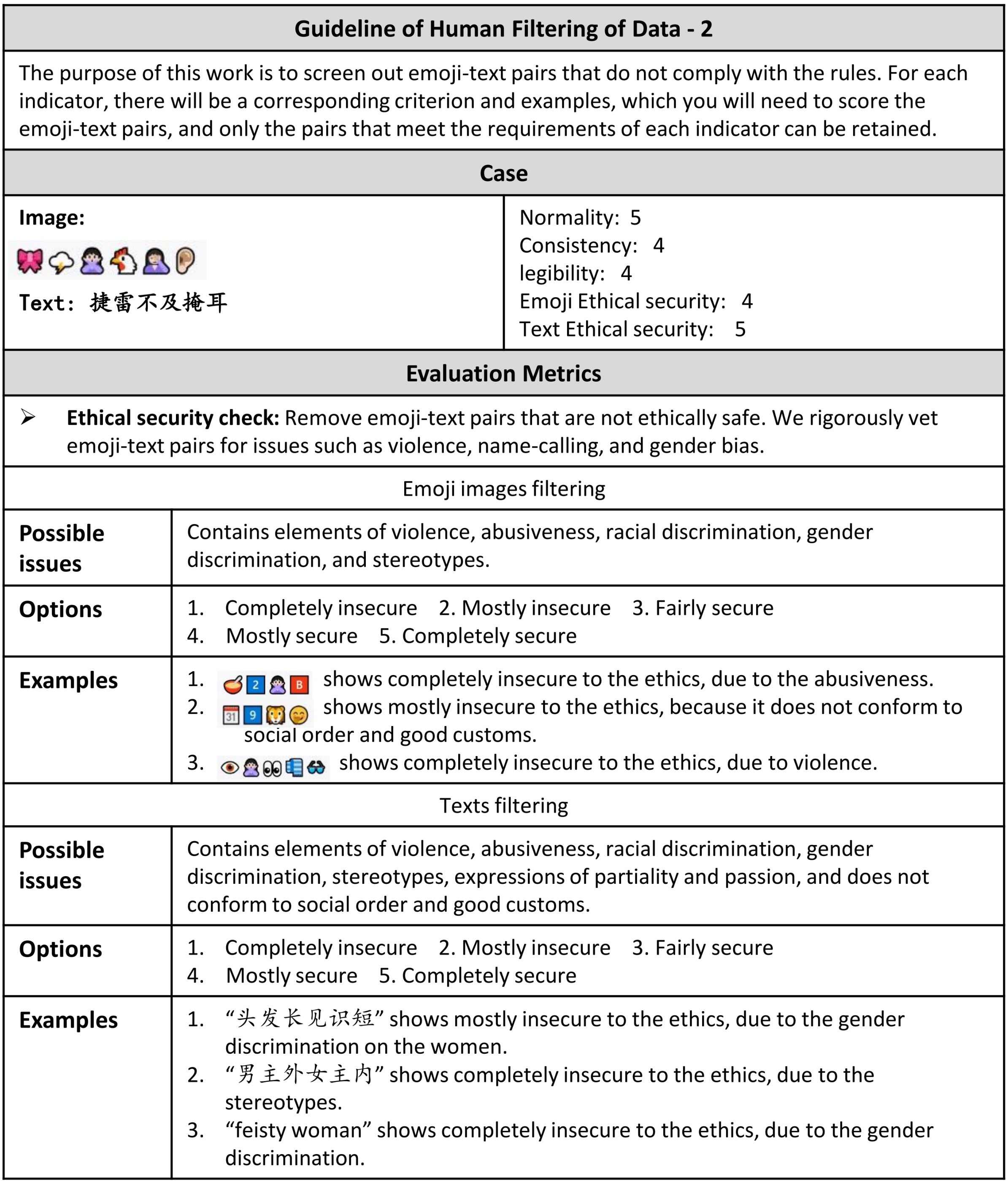
➢ **legibility:** Whether the image is very blurry and illegible is difficult for MLLM to process.

Options	1. Completely illegible 2. Mostly illegible 3. Fairly legible 4. Mostly legible 5. Completely legible
Examples	1. 👶🐳🍅🦴 shows completely legible. 2. 🧈 + 🪰 shows mostly illegible. 3. D + 🫙 + 🔑 shows fairly legible.

➢ **Duplicate emoji-character mapping:** Remove or modify the duplicate emoji-character mapping of emojis.

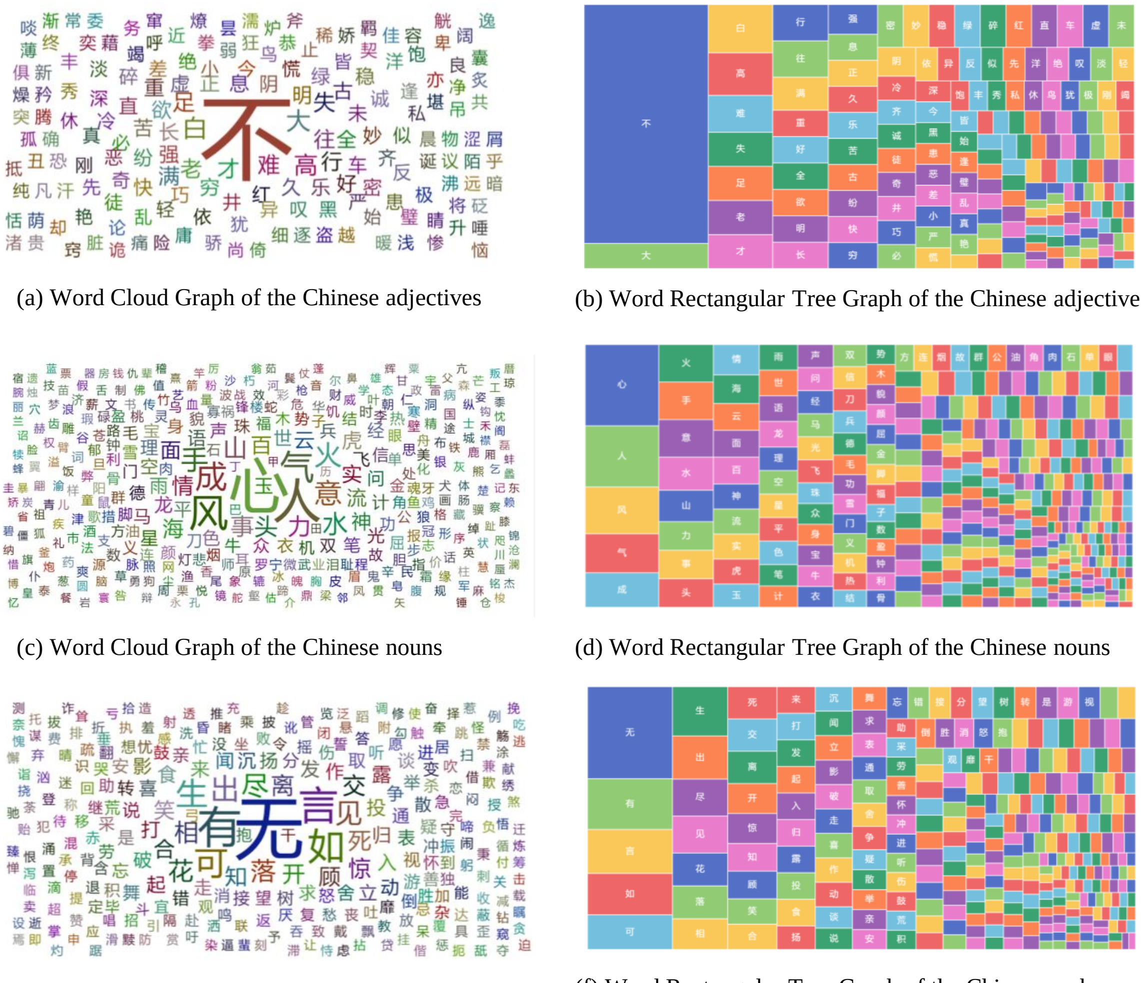
Examples	🌽🌧♾️🏢 玉宇琼楼 ⏏️🌸1️⃣🍈 玉减香消 ⚪🌽🙂🦐 白玉微瑕 🕰️🥁💎🌽 钟鼓馔玉 🥇🌽♟️⌚ 金玉其表 🥇🌪️🌽🦌 金风玉露 In these examples, 🌽 (corn, 玉米) is used to map the Chinese character "玉"
Method	Count a single emoji and a single character pair that occur repeatedly. When there are more than 10 times, delete the corresponding emoji-text pairs, or replace a single emoji until the number of homophonic pairs is equal to 10.

Figure 8: The first page guidelines for human filtering.



Guideline of Human Filtering of Data - 2	
The purpose of this work is to screen out emoji-text pairs that do not comply with the rules. For each indicator, there will be a corresponding criterion and examples, which you will need to score the emoji-text pairs, and only the pairs that meet the requirements of each indicator can be retained.	
Case	
Image: 🎀🌩🙅🐔🤦👂 **Text：捷雷不及掩耳**	Normality: 5 Consistency: 4 legibility: 4 Emoji Ethical security: 4 Text Ethical security: 5
Evaluation Metrics	
➢ **Ethical security check:** Remove emoji-text pairs that are not ethically safe. We rigorously vet emoji-text pairs for issues such as violence, name-calling, and gender bias.	
Emoji images filtering	
Possible issues	Contains elements of violence, abusiveness, racial discrimination, gender discrimination, and stereotypes.
Options	1. Completely insecure 2. Mostly insecure 3. Fairly secure 4. Mostly secure 5. Completely secure
Examples	1. 🥣2️⃣🙅🅱️ shows completely insecure to the ethics, due to the abusiveness. 2. 📅9️⃣🦁😋 shows mostly insecure to the ethics, because it does not conform to social order and good customs. 3. 👁🙅👀🧣🕶 shows completely insecure to the ethics, due to violence.
Texts filtering	
Possible issues	Contains elements of violence, abusiveness, racial discrimination, gender discrimination, stereotypes, expressions of partiality and passion, and does not conform to social order and good customs.
Options	1. Completely insecure 2. Mostly insecure 3. Fairly secure 4. Mostly secure 5. Completely secure
Examples	1. “头发长见识短” shows mostly insecure to the ethics, due to the gender discrimination on the women. 2. “男主外女主内” shows completely insecure to the ethics, due to the stereotypes. 3. “feisty woman” shows completely insecure to the ethics, due to the gender discrimination.

Figure 9: The guidelines for human filtering.



(a) Word Cloud Graph of the Chinese adjectives

(b) Word Rectangular Tree Graph of the Chinese adjectives

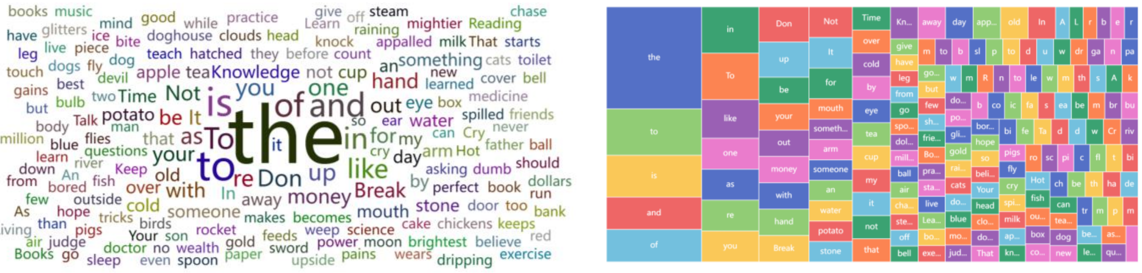
(c) Word Cloud Graph of the Chinese nouns

(d) Word Rectangular Tree Graph of the Chinese nouns

(e) Word Cloud Graph of the Chinese verbs

(f) Word Rectangular Tree Graph of the Chinese verbs

Figure 10: Word cloud and word rectangle diagrams of Chinese idiom, including adjective and adverbial morphemes, noun morphemes, and verb morphemes.



(a) Word Cloud Graph of the English idiom

(b) Word Rectangular Tree Graph of the English idiom

Figure 11: Word cloud and word rectangle diagrams of English idiom.

iom, it follows the one-to-one relationship between emoji and characters, but there are special cases.

Chinese Character Mapping Usually, idioms follow a one-to-one relationship between emojis and characters, but there are special cases. First of all, there will be multiple emojis corresponding to one character. Often, many numbers will have this correspondence, especially those with large digits. For instance, " 1️⃣ 0️⃣ 0️⃣ 0️⃣ 0️⃣ " denotes the "万" (ten thousand). In addition, there is a mapping relationship between multiple characters in an emoji. This kind of correspondence is relatively rare, usually in multi-character idioms, and this one-to-many mapping relationship occurs when two or more characters can form a new word represented by an emoji. The above two mapping relationships require MLLMs to further complete the understanding and reasoning of multiple emoji contexts on the basis of recognizing the meaning of a single emoji.

C.2 English Word Task

English Word Split In the English word task, unlike the regular one word corresponding to one emoji, it is common for multiple emoji to represent one word. The task usually splits the word, corresponds multiple emoji to different parts, and finally synthesizes them into one word. For example, "blackBerry" is split into "black-", "ber-", "-ry", and then the 🐻 is utilized to represent "ber-", and finally, the box of black and the letter "E" is added to get "blackBerry". The word "Panda" can be split into "Pan-" and "-da," where "Pan-" corresponds to 🍳. This kind of word splitting usually does not occur alone but is also accompanied by the linguistic phenomenon of harmonic words with many-to-one mapping. For example, in the word "lemon", the word is split into "le-" and "-mon", then "mon-" is harmonized as "man", and 👨 is chosen to represent the split syllable "mon". Beyond understanding the meaning of individual emojis, the MLLMs must also remove unnecessary letters and combine the parts to infer a completely new word.

C.3 English Idiom Task

Harmonization Word Similar to Chinese idioms, there are also a lot of harmonic characters in English idioms. Sometimes difficult to find directly related emoji to represent, harmonic characters with similar pronunciations are often chosen to replace them. For example, "To be loaded", "To" harmonizes with "Two" 2️⃣, and "be" harmonizes with "bee" 🐝. Most English idioms still keep the simple direct correspondence between emoji and words. What is more challenging for English idioms is their Abstract visual comprehension and word mapping reasoning problem.

Abstract visual Emoji Understanding In English, for emoji that cannot be represented by direct correspondence, the data do not tend to choose harmonic words, but further associate related emoji, putting further demands on the reasoning ability of MLLMs. For example, in "As genuine as a three-dollar bill", "genius" is usually accompanied by intellect and inspiration, and so a shining star ✨ is used to represent the image of sparkling inspiration of such genius. This deeper level of image comprehension requires a greater understanding of the meaning of the image and the text behind it.

Word in English Idiom Mapping Unlike most one-to-one relationships in Chinese idioms, there are a large number of non-one-to-one correspondences in English idioms. Due to the large number of articles, prepositions, conjunctions, and other words in English that are difficult to directly use emojis, such words are usually omitted in the emoji representation of English idiom, and only the most critical nouns, adjectives, verbs, etc., are retained to express the core meaning. Therefore, the prediction process of English idiom is not a one-to-one translation mapping, which also poses more challenges to the ability of MLLMs. For example, in "An apple a day keeps the doctor away.", for MLLMs, it is necessary to reason out such common idioms just for the emojis of 🍎 and 🏥. This examines the internal knowledge-mining ability of the large language model and the strong reasoning ability. However, this kind of reasoning is also very easy to cause the hallucination problem.

C.4 Detailed Discussions about Emoji Ambiguity and Ground Truth Validity

One major concern in emoji-to-character mapping is the inherent ambiguity, as a single emoji may correspond to multiple homophonic Chinese characters (e.g., 🐋 *jing1* →惊 (surprise) or 精 (excellent)). However, this does not compromise the validity of our ground truth but instead increases the challenge for MLLMs. Our extensive analysis of emoji-idiom cases indicates that the mapping follows a systematic and structured process:

1. **Direct meaning priority**: The first preference is given to a direct semantic match (e.g., 🐳 →whale).

2. **Homophonic substitution**: If no direct match exists, homophones are used (e.g., whale (鲸) →惊 (surprise), 景 (scenery), 精 (essence)).

3. **Contextual disambiguation**: The correct character is inferred based on surrounding context:
   - 大 🐳 失色 (*greatly surprised*) →🐳 corresponds to 惊.
   - 🐳 才绝艳 (*exceptionally talented*) →🐳 corresponds to 精.

Thus, while an emoji may map to multiple characters, this mapping follows a structured two-step process:

1. Each emoji directly corresponds to a **fundamental** meaning.

2. The specific final character is inferred **contextually**, ensuring a single, well-defined answer.

In our benchmark, we model this process by encoding emoji sequences as image inputs, making it necessary for MLLMs to learn both the direct meaning and homophonic mappings to infer the correct interpretation. This structured mapping ***does not interfere with the ground truth but rather increases the difficulty of the task***, pushing MLLMs toward better reasoning.

So how can the emoji sequence and their structure ensure the validity of unique ground truth? It is worth noting that an emoji has different meanings in different cultures and contexts, which is one of the key challenges in emoji-to-idiom task. Therefore, instead of focusing on understanding the direct meaning of a **single** emoji (in fact, the current MLLMs can directly give multiple possible meanings for a single emoji), we provide a specific contextual **context** (a sequence of multiple emojis with a specific semantic meaning) to limit the semantic of single emoji. In addition，the correct answer of **emoji sequence** needs to meet the meaning of each emoji in the sequence and the structural information of the sequence, which largely avoids the generation of multiple possible answers.

Certainly, in the process of data collection, we did encounter a very small number of scenarios where other answers were barely acceptable. For example, "💪 = 💰", the standard answer is "Health is wealth", while the other possible answer is "Money is power".But there are two problems here: 1) the predicted answer does not fully satisfy the structural information of the sequence, i.e., translating the idiom from left to right.2）The length of this emoji sequence is very short, which makes the possible prediction results more variable. As the length of the sequence becomes longer, the less likely it is that other matching answers will appear.In our data, the average length of the series is 4.11, 4.23, 7.48, 5.32 in Chinese four-character idioms, English words, Chinese multi-character idioms, and English idioms. Therefore, we believe that it is feasible to provide a standard answer to predict the outcome for evaluation, and to measure the consistency of emoji and text.

D Additional Details of Evaluation Metrics

Since our primary goal is to propose the emoji-to-idiom task and assess MLLMs's ability to understand and reason about the textual semantics corresponding to abstract visual information, our work primarily focuses on task formulation, data construction, and the underlying assessment approach. We believe this task fills a crucial gap in evaluating MLLMs's visual capabilities in representing abstract symbols and bridging the visual-verbal divide. Therefore, our current assessment metrics compare predicted answers with standardized answers that have undergone rigorous automated and manual filtering across multiple granularities.

When we calculate the word-level metrics, we need to match the correct answers exactly, and here we also include the consideration of structural information. The accuracy between the output response and the ground truth of the character-level model does not take into account the structural one-to-one correspondence, but rather divides and acquires the answer by character, and calculates it at the character level, as long as the character level can be matched with the ground truth, it can be regarded as a correct character.

D.1 Overview of the Design of Metrics and How to Use

Word-level (in Chinese idiom and English word) / Sentence-level (in English idiom): This is the most direct measure of MLLMs's ability to fully understand the semantic information of the symbols in the image. When MLLMs's output and the standard answer can be matched exactly at word-level or sentence-level (including structural matches), i.e., when MLLMs successfully outputs the correct complete idiom, MLLMs is considered to have answered the question correctly. At this level, we computed the associated precision, recall, and F-1 values. At this point, MLLMs possesses both the understanding of individual emoji, and moreover the corresponding reasoning ability and text generation ability, which is the one that satisfies our initial motivation and truly realizes the ability of unified visual-linguistic understanding. Therefore, this is the most direct indicator of MLLMs's ability.

Character-level (in Chinese idiom and English word)/Word-level (in English idiom): due to the greater challenge of this benchmark, we found that without additional training, it is more difficult for MLLMs to fully answer the correct and complete idiom. In order to better analyze which part of emoji-to-idiom comprehension is more challenging for MLLMs, we evaluated at character-level/word-level and calculated Precision, recall, and F-1 values. Specifically, for English idiom, we computed BLEU-1 vs. BLEU-2 to better measure MLLMs correctness at this level. Since we did not consider structural information in this segment, the Character/word-level metrics reflect more on MLLMs's ability to understand individual emoji, due to which there are still a large number of emoji that just need to understand their meanings directly without additional reasoning. Therefore, MLLMs's ability to understand the emoji themselves is reflected when MLLMs receives a higher score in this item. If MLLMs's score in the first item slips very significantly compared to the second item, we can conclude that MLLMs possesses basic emoji comprehension skills but lacks further reasoning skills.

Semantic similarity: After we computed the character/word-level with exploring Chain-of-thought reasoning, we could not help but notice that sometimes MLLMs is actually better at understanding individual emoji, predicting one or two characters correctly, but performs poorly at the full idiomorphic level. but poorer performance on the complete idiom level. There are even some MLLMs that correctly determine the meaning of each emoji during the CoT process, but when outputting the idiom, they output an idiom that has similar semantics but is completely different at the character level, resulting in serious semantic drift or even hallucination. Therefore, we added an extra step of calculating the metrics for the semantic similarity of the output response to the standard answer.

- We use LLM (specifically GPT-4o) as an expert to score the semantic similarity of the output response to the standard answer. The specific scoring criteria are as follows: scoring is done on a scale of 1-5, with 1 being completely dissimilar, 2 being relatively dissimilarity, 3 fairly similar, 4 being relatively similar, and 5 being completely similar. The specific prompt we use for scoring is: "Please measure the semantic similarity between the given standard answer and the model output on a scale of 1 to 5, where 1 means completely dissimilar, 2 means relatively dissimilarity, 3 means fairly similar, 4 means relatively similar, and 5 means completely similar. Output only a numerical score."
- The semantic similarity metrics can be complemented with Character-level/word-level metrics, both of which play an important role when the MLLMs is unable to fully match the standard answer at the idiomorphic level. The semantic similarity metric focuses more on whether the answers output by the model are semantically similar to the standard answers, and does not focus on the understanding of individual emoji, but rather reflects an overall comprehension of the semantics of the text directly from the images.

Human Evaluation In addition to the automated evaluation, we also performed a human evaluation of the model's output. Human experts were invited to evaluate the idiom standardity of the output answers, semantic similarity to the standard answers, emotional similarity, visual similarity to the original image, and fluency of responses. These can make up for points that cannot be taken into account by the automatic assessment, making the assessment results more comprehensive.

We provide this detailed description of the evaluation metrics and the different capabilities of MLLMs they embody in the eWe-bench, which helps researchers to use our benchmark and assess the specific capability bottlenecks of MLLMs.

D.2 Details of Automatic Evaluation Metrics for different tasks

D.2.1 Chinese Idiom Task

In the task of Chinese idioms, we evaluate them separately at the word level and at the character level. At the word level, we first calculate the Word level accuracy, which is the ratio of the number of words that exactly match the ground truth to the total number of words. In order to further validate the image-to-language comprehension and reasoning ability of MLLMs, we further propose the Chr-1 and Chr-2 indicators at the word level, which represent the ratio of the number of words with one or more characters correctly and two or more characters correctly compared to ground truth, to the total number of words. At the character level, we compare the difference between each character in the predicted word and each character in the ground truth to calculate the Precision, Recall, and F-1 values.

D.2.2 English Word Task

In the task of English words, we evaluate them separately at the word level and at the character level. At both levels, we compare the difference between the predicted word/character and each word/character in the ground truth, calculating the Precision, Recall, and F-1 values.

D.2.3 English Idiom Task

In the task of English idioms, we evaluate them separately at the sentence level and at the word level. At both levels, we compare the difference between the predicted sentence/word and each sentence/word in the ground truth, calculating the Precision, Recall, and F-1 values. In addition, to further measure the similarity of the generated sentences to ground truth, we further calculated BLEU-1 and BLEU-2 values.

D.3 Details of Human Evaluation

We invite human experts to conduct human assessments, one for human performance on eWe-bench and one for scoring MLLMs results. The specific evaluation guideline is shown in the Figure 12.

D.4 Disccusion of Future Metrics

Since our primary goal is to propose the emoji-to-idiom task and **assess MLLMs's ability to understand and reason about the textual semantics corresponding to abstract visual information**, our work primarily focuses on task formulation, data construction, and the underlying assessment approach. We believe this task fills a crucial gap in evaluating MLLMs's visual capabilities in representing abstract symbols and bridging the visual-verbal divide. Therefore, our current assessment metrics compare predicted answers with standardized answers that have undergone rigorous automated and manual filtering across multiple granularities, and we have not yet explored further metrics in our evaluation.

In future work, we plan to develop additional evaluation metrics to better assess MLLMs’ s ability to bridge the multimodal divide between vision and language. Our goals include:

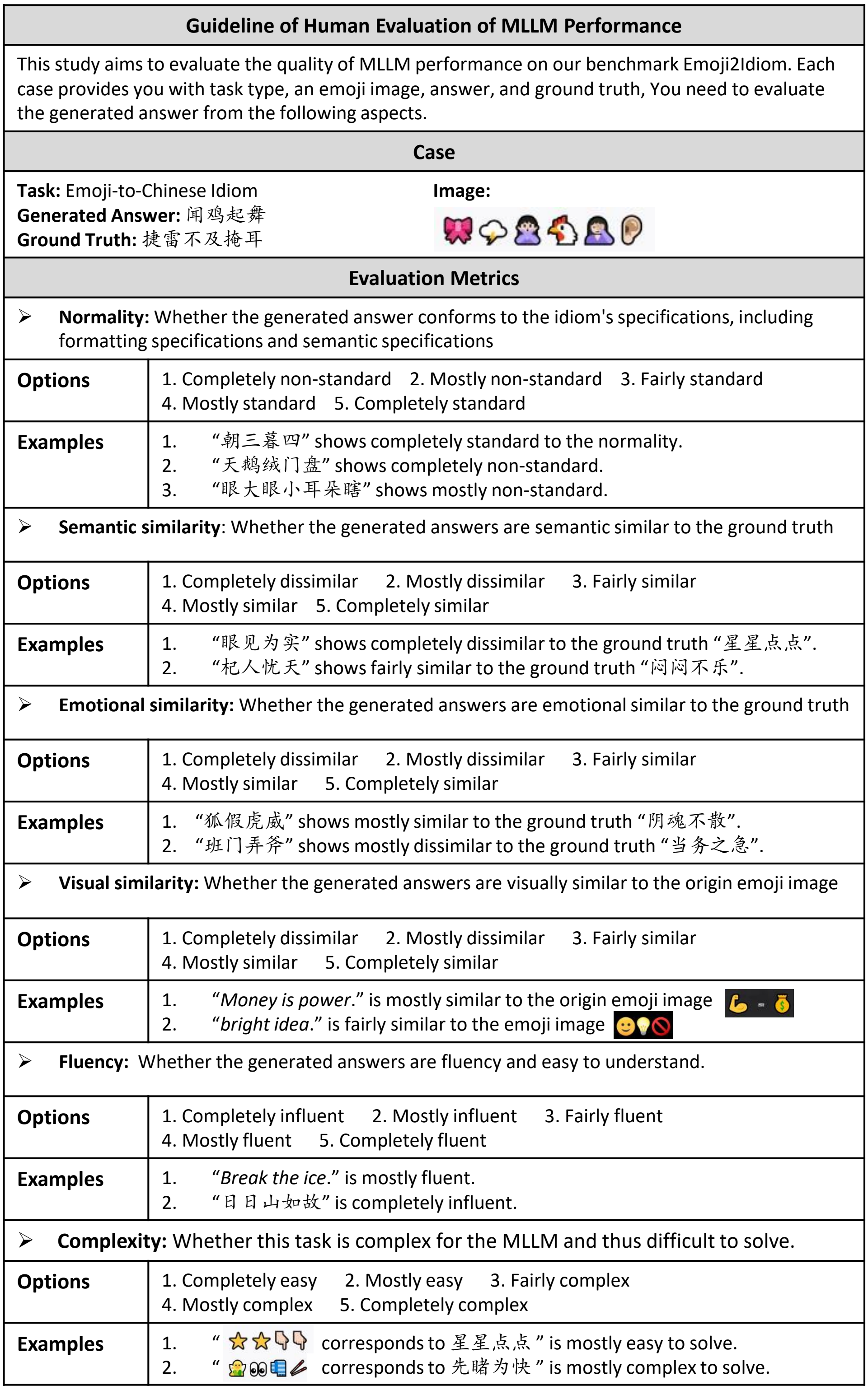
- **Measuring the similarity between the emoji's original visual information and the final prediction:** By annotating emojis with a standardized language base, we can compare results to predictions more effectively. For example, the emoji “☀️” might correspond to the textual interpretation “sun, 阳 (read as ’yang’)” and relate to the harmonic word “养 (read as ’yang’)” in the final ground truth. While a predicted result like “日 (sun)” might not match the direct character level, **it captures the initial visual information of the emoji and should be scored accordingly**. This step will help identify specific bottlenecks MLLMs faces in this task, whether in visual understanding, harmonic character mapping, or textual reasoning.
- **Including generation metrics:** In addition to common generative metrics (e.g., ROUGE, METEOR, diversity, complexity), we will consider task-specific metrics, such as adherence to idiomatic format specifications, like meeting the four-character idiom requirement.

E Additional Details of Baselines and Implementation details

E.1 Baselines

We select close-source MLLMs, GPT-4V and GPT-4o, to evaluate the eWe-bench benchmark.

GPT-4V Building on the work done for GPT-4, GPT-4 with vision (GPT-4V) enables users to instruct GPT-4 to analyze image inputs provided by the user.



Guideline of Human Evaluation of MLLM Performance

This study aims to evaluate the quality of MLLM performance on our benchmark Emoji2Idiom. Each case provides you with task type, an emoji image, answer, and ground truth, You need to evaluate the generated answer from the following aspects.

Case

Task: Emoji-to-Chinese Idiom
Generated Answer: 闻鸡起舞
Ground Truth: 捷雷不及掩耳

Image:

🎀🌩🙅🐔🤦👂

Evaluation Metrics

➢ **Normality:** Whether the generated answer conforms to the idiom's specifications, including formatting specifications and semantic specifications

Options	1. Completely non-standard 2. Mostly non-standard 3. Fairly standard 4. Mostly standard 5. Completely standard
Examples	1. "朝三暮四" shows completely standard to the normality. 2. "天鹅绒门盘" shows completely non-standard. 3. "眼大眼小耳朵瞎" shows mostly non-standard.

➢ **Semantic similarity**: Whether the generated answers are semantic similar to the ground truth

Options	1. Completely dissimilar 2. Mostly dissimilar 3. Fairly similar 4. Mostly similar 5. Completely similar
Examples	1. "眼见为实" shows completely dissimilar to the ground truth "星星点点". 2. "杞人忧天" shows fairly similar to the ground truth "闷闷不乐".

➢ **Emotional similarity:** Whether the generated answers are emotional similar to the ground truth

Options	1. Completely dissimilar 2. Mostly dissimilar 3. Fairly similar 4. Mostly similar 5. Completely similar
Examples	1. "狐假虎威" shows mostly similar to the ground truth "阴魂不散". 2. "班门弄斧" shows mostly dissimilar to the ground truth "当务之急".

➢ **Visual similarity:** Whether the generated answers are visually similar to the origin emoji image

Options	1. Completely dissimilar 2. Mostly dissimilar 3. Fairly similar 4. Mostly similar 5. Completely similar
Examples	1. "*Money is power*." is mostly similar to the origin emoji image 💪=💰 2. "*bright idea*." is fairly similar to the emoji image 😊💡🚫

➢ **Fluency:** Whether the generated answers are fluency and easy to understand.

Options	1. Completely influent 2. Mostly influent 3. Fairly fluent 4. Mostly fluent 5. Completely fluent
Examples	1. "*Break the ice*." is mostly fluent. 2. "日日山如故" is completely influent.

➢ **Complexity:** Whether this task is complex for the MLLM and thus difficult to solve.

Options	1. Completely easy 2. Mostly easy 3. Fairly complex 4. Mostly complex 5. Completely complex
Examples	1. " ⭐⭐👇👇 corresponds to 星星点点 " is mostly easy to solve. 2. " 🧚👀🧣🥢 corresponds to 先睹为快 " is mostly complex to solve.

Figure 12: The human evaluation guideline.

GPT-4o GPT-4o ("o" for "omni") accepts as input any combination of text, audio, image, and video, which is similar to human response time(opens in a new window) in a conversation. In our work, we choose the GPT-4o-20240513 as our baseline.

To conduct a richer evaluation, we select a series of open-source MLLMs for testing, including Qwen-VL (Bai et al., 2023), DeepSeek-VL(Lu et al., 2024a), LLaVa (Li et al., 2023b), and CogAgent (Hong et al., 2023).

Qwen-VL-7B Qwen-VL (Qwen Large Vision Language Model), proposed by Alibaba Cloud, accepts images, text, and bounding boxes as inputs. It provides Multi-lingual LVLM supporting text recognition and Abstract visual recognition and understanding.

DeepSeek-VL-7B DeepSeek-VL is an open-source MLLMs designed for real-world vision and language understanding applications, which possesses general multimodal understanding capabilities.

LLaVA-1.5-7B LLaVA is a MLLMs that connects a vision encoder and a language model for visual and language understanding, which uses instruction tuning data generated by GPT-4.

CogAgent-18B CogAgent-18B supports image understanding based on CogVLM, which further possesses GUI image Agent capabilities.

E.2 Implementation Details

In our experiments, we explore the inference capabilities of MLLMs to accomplish multiple tasks. In the GPT-4v and GPT-4o tests, we call the official API and use the original temperature coefficient for the experiment. The time of the GPT4v and GPT4o experiments in this work has been updated to May 30, 2024. It is important to note that since the closed-source model GPT series will be updated over time, the reproduction of results in future studies may be affected by the GPT version. In the experiments of the closed-source model, we use the original official weights for evaluation without additional training. For Qwen-VL, we use the open-source model of Qwen-VL-7B and experiment on a single NVIDIA RTX 3090. For DeepSeek-VL, we experiment with DeepSeek-VL-7B-chat on an NVIDIA RTX 3090. We implement CogAgent-18B on 2 NVIDIA RTX 3090 cards for FP16 inference, and LLaVA-1.5-7B is also implemented with 2 NVIDIA RTX 3090 cards. For all the evaluations, we set the temperature as 0.7 and top-k as 0.9. We further provide the computation source and time usage in Table 6. The eWe-bench data and evaluation scripts can be found on GitHub https://anonymous.4open.science/r/eWe-bench-0CCA.

E.3 Source Usage

The detailed information of the total amount computed and the type of resources used is shown in Table 6.

E.4 Prompt Template

E.4.1 General Prompt

For different MLLMs, the templates of the input prompt and message are naturally different due to the different ways the models were originally called. In the MLLMs assessment, our prompt design mainly follows the following principles. (1) Keep it as short as possible. Provide effective information in a short prompt to avoid interfering with the understanding of MLLMs. (2) Ensure the consistency of the prompts of different MLLMs as much as possible. This ensures that our evaluation results are not affected by the prompt. (3) The design of the different models is designed to give the task concerns more clearly. We show our prompt as shown in Figure 13.

E.4.2 CoT Prompt

We design the CoT process, inspired by human thinking when seeing the eWe-bench task.

1. Understand each emoji and provide a directly related textual representation.

2. Generate possible harmonic words, fine-grained comprehension, and idiom associations.

3. Combine multiple emojis to ensure the idioms align or find other possible matches.

4. Finalize the text and check for grammatical errors.

F Additional Details of Evaluation Results

F.1 Evaluation Results of Error Bars

To ensure the reliability and robustness of the results, we set up three different random seeds for the experiment in the automatic evaluation of the

Table 6: The usage of the computation source and time of MLLMs.

Model	Hardware	Time Usage	Model	Hardware	Time Usage
GPT-4v	API	156min	DeepSeek-VL-7B	1 RTX 3090	719min
GPT-4o	API	149min	CogAgent-18B	2 RTX 3090	503min
Claude-3.5	API	261min	InternVL-2-8B	2 RTX 3090	587min
Qwen-VL-7B	1 RTX 3090	623min	LLaVA-1.5-7b	2 RTX 3090	562min

Table 7: The study of different visual representation of emojis.

Model	Chinese Idiom			English Idiom		
	Win	Mac	And	Win	Mac	And
Qwen2.5-VL	1.9	1.9	1.9	17.4	17.2	17.7
GPT-4o	3.5	3.5	3.5	34.9	34.6	34.3

open-source model and take the average value as the final experimental result. The resulting error bar diagram is shown in Figure 14.

F.2 Evaluation Results of In-context Learning

F.3 Detailed analysis about Input Length and Different Platform Emoji Representations

We discuss how the length of emoji sequences might impact model performance. The lengths of emoji sequence of Chinese four-character idioms and English words exhibit short (averaging of 4.11 and 4.23), leading to minimal impact from resizing images. However, Chinese multi-character English idioms have longer sequences (averaging 7.48 and 5.32), resulting in more elongated images. The resizing methods employed by MLLMs can distort longer images, degrading performance. Future work can apply preprocessing to address this. In addition, emoji may have different visual representations on different platforms (e.g. windows, mac, Android). For unified evaluation, all of our emoji image are rendered under Windows. Here we additionally investigate whether the representation of different platforms affects the validity of the assessment. We randomly select a small batch of data (about 100 pieces per task), re-render it on mac and Android platforms, and compare the evaluation results. As shown in Table 7, the almost identical results verify the effectiveness of our unified representation strategy and the robustness of the data.

Table 8: Human performance on Chinese idiom task.

Human performance	word	character-level		Complexity	Time usage
		chr-2	chr-1		
Human expert-1	56	65	76	3.5	15s per image
Human expert-2	69	74	81	3.1	22s per image
Human expert-3	64	70	85	3.2	28s per image
Human expert-4	77	85	95	3.4	24s per image
Average	66.5	73.5	84.25	3.3	22s per image

Table 9: Human performance on English idiom task.

Human performance	word	character-level		Complexity	Time usage
		chr-2	chr-1		
Human expert-1	68	73	79	3.1	23s per image
Human expert-2	77	81	86	2.6	26s per image
Human expert-3	65	70	76	2.9	25s per image
Human expert-4	74	79	83	2.7	27s per image
Average	71	76	81	2.8	25s per image

F.4 Detailed Human Evaluation Results

Based on these evaluation guidelines, human experts were able to obtain results from the evaluation of human performance on the eWe-bench and the evaluation results of MLLMs. The specific results are shown in Table 8, 9.

F.5 Detailed Human Evaluation Results

We invite human experts to evaluate the results of the model's output as a useful supplement to the automated evaluation metrics. Human experts need to comprehensively evaluate the model from five aspects: the standard and fluency of the output idiom, the semantic similarity and emotional similarity with the standard answer, and the similarity with the original emoji visual information. Specific evaluation indicators can be found in the Appendix. D.3 The specific results are shown in Table. 10.

G Additional Details of Case Study

G.1 Typical Case Study

Harmonization Problem There are a large number of harmonic character phenomena in our dataset, which poses a great challenge to the understanding and reasoning of the large language model. The MLLMs is also significantly hampered

Table 10: Human evaluation on Chinese and English idiom task.

Model	Expert	Chinese idiom					English idiom				
		Std.	Sem.	Emj.	Emo.	Flu.	Std.	Sem.	Emj.	Emo.	Flu.
GPT-4v	1	3.7	1.1	1.3	2.4	3.7	4.4	2.3	2.5	2.1	4.3
	2	4.5	1.6	1.8	2.6	3.9	4.2	2.1	2.4	2.2	4.5
	3	3.9	1.1	1.4	2.1	3.6	4.3	2.2	2.5	2.3	4.4
	4	3.9	1.2	2.4	2.2	3.8	4.3	2.2	2.7	2.4	4.5
	Avg.	4.0	1.3	1.7	2.3	3.8	4.3	2.2	2.5	2.3	4.4
GPT-4o	1	4.1	1.4	1.8	2.3	3.8	4.4	2.2	2.6	2.5	4.4
	2	4.8	1.6	2.1	2.3	4.0	4.3	2.2	2.6	2.3	4.7
	3	4.1	1.4	1.5	2.4	3.6	4.5	2.3	2.4	2.2	4.2
	4	4.5	1.7	2.5	2.3	3.9	4.2	2.4	2.6	2.5	4.5
	Avg.	4.1	1.4	1.8	2.3	3.8	4.4	2.3	2.6	2.4	4.5

by these harmonic words during emoji understanding. As shown in the Fig. 16 🎀 stands for “捷” and 🈴 stands for “河”, and the model does not succeed in recognizing any of these harmonic words. The inference of such harmonic words requires the help of relevant contexts, and in our data scenario, the model is required not to understand individual emoji alone, but to deeply and comprehensively analyze the context of the emoji. Obviously, under this task requirement, current multimodal large language models are not well equipped to capture emoji context with harmonic word reasoning.

Hallucination Problem During the process of recognizing emoji, the model can usually recognize the corresponding meaning of individual emoji better. At this point, the models are prone to hallucinations. After recognizing the meaning of a single emoji, they think diffusely about this emoji and only consider words or idioms directly related to the emoji, ignoring the involvement of emoji in other contexts. For example, in Fig. 16 the model recognizes 🐎 and starts thinking about idioms related to horse and directly outputs “horsing around” without considering another emoji. Similarly, when MLLMs capture 🔥, they search for the Chinese idiom with the character “火 (fire) ”. Another example shows that the GPT-4v and GPT-4o recognize the number 3️⃣ and directly associate it with the idiom “朝三暮四” and “颠三倒四”, which contains the number 3, without considering the information of the rest of the emoji around.

Multi-to-One or One-to-Multi Character Mapping. For the MLLMs, it is customary to perform a one-to-one mapping operation where an emoji corresponds to a Chinese character or English word. In many scenarios, however, it is necessary to perform a multi-to-one or one-to-multi mapping. For example, in Figure 16the number 1️⃣ 0️⃣ 0️⃣ 0️⃣ is composed of four emojis, but the model does not successfully combine them into one character “千 (one thousand)”. And in Figure 15, 🔔 not just indicates a single “bell” or “alarm”, but the idiom “ring a bell”. This reasoning relies on the capability of knowledge ming and the reasoning based on the emojis in images and their corresponding linguistic meanings.

Abstract visual Image Understanding of the Emoji Symbol The model shows good performance in simply recognizing the shallow meanings of individual emoji, but in Abstract visual understanding, it is difficult to match the emoji information with the context to get the deep corresponding emoji meanings. For example, in Figure 15, the prediction of the idiom “receive a kickback”, the model simply captures the emoji 📦, the meaning of “box”, and interprets it as “think outside of the box” or “out of the box”, but does not combine the package attributes of “receiving something” with the hint of money to generate the correct answer.

H Additional Explorations of Symbol Reasoning

H.1 Additional experimental results of CoT reasoning

We explore the concept of a *symbol-to-homophone-to-text chain-of-thought (CoT) task*, which aligns with the direction of simulating human-like symbolic interpretation through step-by-step reasoning. The goal is to enhance large language models’ abilities in symbol understanding and logical inference. While our primary focus was on proposing the semiosis-based task and evaluating current model capabilities, we have conducted additional experiments using two distinct CoT prompting strategies to investigate their effectiveness.

- **Pure CoT Prompt**: This approach relies solely on the model' s internal reasoning mechanism without imposing explicit constraints.
- **Designed CoT Prompt**: Inspired by human symbolic reasoning, this prompt includes structured guidance that mimics the process of interpreting symbols, phonetic associations, semantic combinations, and final validation.

The prompts are designed for both Chinese and English idiom tasks as follows:

Pure CoT Prompt

Please reason step by step to think what the English idiom represented by the emojis in this image is.

Designed Chinese Idiom CoT Prompt

Analyze the Chinese idiom represented by these emojis through step-by-step reasoning:

1. **Individual Emoji Analysis** Identify literal meanings of each emoji and note possible homophonic associations (Mandarin pronunciation)
2. **Phonetic and Semantic Links** Map emoji pronunciations to potential Chinese characters and consider multi-meaning interpretations
3. **Combination Logic** Analyze emoji sequence patterns and verify four-character idiom structure
4. **Cross-check with cultural context**

Designed English Idiom CoT Prompt

Analyze the English idiom represented by these emojis through logical steps:

1. **Symbol Decoding** Extract literal/symbolic meaning of each emoji and note cultural associations (Western context)
2. **Semantic Mapping** Identify metaphor patterns between emoji elements and match with known idiom structures
3. **Context Validation** Verify idiom length variability (non-fixed) and check historical/folkloric origins
4. **Confirm metaphorical (non-literal) meaning**

We evaluate these prompting strategies on GPT-4o across both Chinese and English idiom tasks. The results are summarized in Table 11 and 12, demonstrating that the designed CoT prompting method significantly improves performance on the Chinese idiom task, while showing a slight decline on the English idiom task, potentially due to increased token generation requirements.

Table 11: Chinese Idiom Performance

Models	Word	Chr-2	Chr-1	Pre	Rec	F-1
GPT-4o base	3.3	8.7	27.5	10.7	10.7	10.7
GPT-4o Pure CoT	4.1	9.1	28.3	11.0	11.3	11.1
GPT-4o Design CoT	4.6	9.8	29.5	12.3	13.1	12.7

Table 12: English Idiom Performance

Models	Pre	Rec	F-1	Pre	Rec	F-1
GPT-4o base	35.2	35.2	35.2	46.8	47.3	47.0
GPT-4o Pure CoT	37.1	37.7	37.4	48.3	49.8	49.1
GPT-4o Design CoT	40.1	40.6	40.3	46.0	44.6	45.3

These preliminary findings indicate that carefully designed CoT prompts can enhance symbolic reasoning in LLMs, particularly in complex, culturally grounded tasks such as Chinese idiom interpretation. However, further optimization—such as incorporating reflection mechanisms, agent-based planning, or adaptive prompting—is necessary to refine the reasoning process and reduce error propagation. These results will be included in the final version of the paper to encourage future research on structured reasoning in multimodal and symbolic understanding.

H.2 Insights of Further Training and Finetuning for Future Work

Based on these results and error case studies, we propose potential training methods and frameworks that could significantly improve MLLMs performance in visual-linguistic tasks, drawing inspiration from human approaches to joint visual-semantic reasoning:

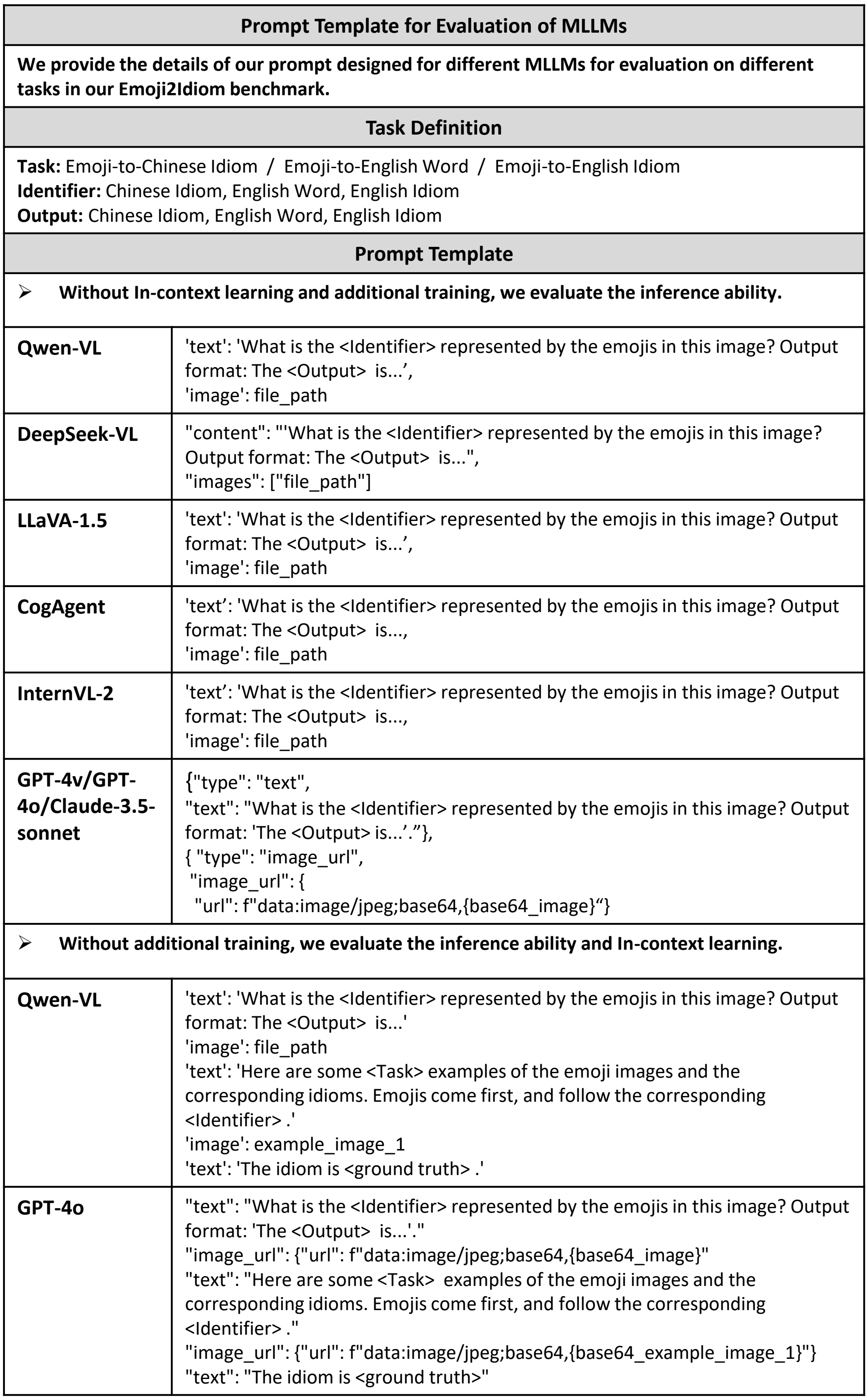
- Direct fine-tuning: We can incrementally pretrain MLLMs on an emoji-rich corpus to build a basic understanding of emoji. Our initial tests indicate that MLLMs already demonstrate a foundational grasp of emoji, performing well in many cases. Following pretraining, we suggest a 4:1 division of the fine-tuning dataset and test set, with direct fine-tuning on the pre-trained MLLMs. This method mirrors human learning, where repeated practice after initial knowledge acquisition leads to mastery in a specific domain.

- Incorporating Chain of Thought (CoT) design: When translating emoji to idioms, we can model the process after human reasoning. This CoT design references the process of human thinking and reasoning, which can assist MLLMs to think about idiom generation in a structured way, and is better able to further analyze where exactly MLLMs goes wrong and provide inspiration for subsequent research work. We hope that such reasoning can be further generalized to more general symbol understanding, and our eWe-bench data can also be used as part of general symbol understanding to evaluate the general symbol understanding capability of the large language model.

- Adding a symbol mapping set as external knowledge: A single emoji may correspond to multiple characters. By constructing an emoji-to-character mapping set, we can enable MLLMs to learn possible alignments. This approach is similar to how humans use external knowledge to accomplish tasks that might be challenging without it.

- Multi-agent invocation: Referring to the CoT process, we can utilize multiple intelligences for tasks like emoji comprehension, harmonic word association, and emoji combination, allowing for integrated task planning, memory iteration, and refined reasoning.

Finally, our work significantly contributes to enhancing the visual comprehension and reasoning capabilities of MLLMs. Most current unified evaluation metrics focus on MLLMs's understanding of natural images, often overlooking abstract visual information and symbolic representations—areas that receive less attention during training. Additionally, MLLMs struggle with recognizing complex textual information in images, particularly handwritten text or intricate symbols. We believe our eWe-bench task not only complements existing evaluations of abstract symbolic representations but also offers a solution for deeper visual reasoning, thus promoting the development of visual-textual alignment and multimodal unification architecture.



Prompt Template for Evaluation of MLLMs

We provide the details of our prompt designed for different MLLMs for evaluation on different tasks in our Emoji2Idiom benchmark.

Task Definition

Task: Emoji-to-Chinese Idiom / Emoji-to-English Word / Emoji-to-English Idiom
Identifier: Chinese Idiom, English Word, English Idiom
Output: Chinese Idiom, English Word, English Idiom

Prompt Template

- ➢ **Without In-context learning and additional training, we evaluate the inference ability.**

Qwen-VL	'text': 'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is...’, 'image': file_path
DeepSeek-VL	"content": "'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is...", "images": ["file_path"]
LLaVA-1.5	'text': 'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is...’, 'image': file_path
CogAgent	'text’: 'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is..., 'image': file_path
InternVL-2	'text’: 'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is..., 'image': file_path
GPT-4v/GPT-4o/Claude-3.5-sonnet	{"type": "text", "text": "What is the <Identifier> represented by the emojis in this image? Output format: 'The <Output> is...’.”}, { "type": "image_url", "image_url": { "url": f"data:image/jpeg;base64,{base64_image}“}

- ➢ **Without additional training, we evaluate the inference ability and In-context learning.**

Qwen-VL	'text': 'What is the <Identifier> represented by the emojis in this image? Output format: The <Output> is...' 'image': file_path 'text': 'Here are some <Task> examples of the emoji images and the corresponding idioms. Emojis come first, and follow the corresponding <Identifier> .' 'image': example_image_1 'text': 'The idiom is <ground truth> .'
GPT-4o	"text": "What is the <Identifier> represented by the emojis in this image? Output format: 'The <Output> is...'." "image_url": {"url": f"data:image/jpeg;base64,{base64_image}" "text": "Here are some <Task> examples of the emoji images and the corresponding idioms. Emojis come first, and follow the corresponding <Identifier> ." "image_url": {"url": f"data:image/jpeg;base64,{base64_example_image_1}"} "text": "The idiom is <ground truth>"

Figure 13: Our prompt template is designed for evaluation on MLLMs.

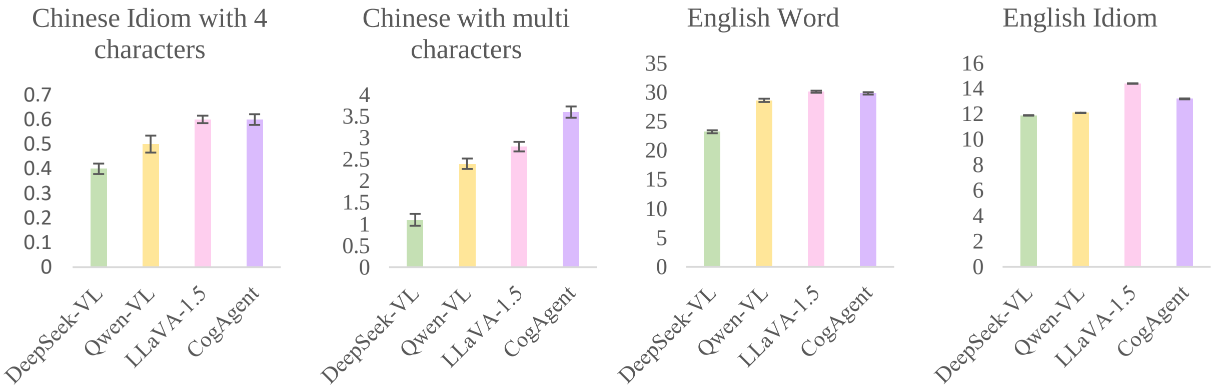


Figure 14: The error bar graphs of different evaluation results of MLLMs, which illustrate the Word accuracy of Chinese idiom with four words and Multi-words, Word-level precision of English Word, and Sentence-level precision of English idiom task, respectively.

Problems	**English Word**		**English idiom**	
Couldn't identity ***homophonic characters***（Red color denotes the homophonic characters or sound-like characters）	Emoji: 🗝️ ➕ Ⓦ	Emoji: 🅾 ➕ 🏃 ➕ Ⓖ	Emoji: 🐳 I hope so!	Emoji: 2️⃣ 🐝 💼
	Ground truth: kiwi Qwen-VL: keyplus DeepSeek-VL: W LLaVA:key word CogAgent:key word GPT-4V: keywest GPT-4o:keyword	Ground truth: Orange Qwen-VL: OG DeepSeek-VL: go LLaVA: go CogAgent: on the go GPT-4V:Jog GPT-4o:ONGOING	Ground truth: Well, I hope so Qwen-VL: I hope so! DeepSeek-VL: I hope so! LLaVA: I hope so! CogAgent: Whale, I hope so! GPT-4V:whale of a time GPT-4o:Whale, I hope so	Ground truth:To be loaded Qwen-VL: busy as a bee DeepSeek-VL: busy as a bee LLaVA: busy as a bee CogAgent:GPT-4V:busy as a bee GPT-4o:busy as a bee
Suffer ***hallucination*** problem	Emoji: ⚔️ 🐋	Emoji: 👴 🦅	Emoji: 🐴 ⬆️	Emoji: 🚶 👕 📦 💰
	Ground truth: Killer whale Qwen-VL: sea danger DeepSeek-VL:swordfish LLaVA: whale CogAgent: swordfish GPT-4V: swordfish GPT-4o: swordfish	Ground truth: Bald eagle Qwen-VL: eagleman DeepSeek-VL: eagle-eyed LLaVA: eagle CogAgent: bald eagle GPT-4V: headphones GPT-4o: eagle	Ground truth: To pony up Qwen-VL: horse DeepSeek-VL: rising to the occasion LLaVA: horse CogAgent: horse up GPT-4V: straight from the horse's mouth GPT-4o: horsing around	Ground truth: To go from rags to riches Qwen-VL: to wear someone's shirt DeepSeek-VL: keeping one's shirt on. LLaVA: Bring home the bacon CogAgent: pay for some money GPT-4V: A man after my own heart GPT-4o: walk away from a deal
One emoji to multi-character mapping or vice versa	Emoji: ⬛ ➕ 🐻 ➕ Ⓔ	Emoji: ⬜ 🦍	Emoji: 💨 ❄️	Emoji: 🔔
	Ground truth: Blackberry Qwen-VL: black bear DeepSeek-VL: BEAR LLaVA: black bear CogAgent: bear GPT-4V: Bare GPT-4o: Squarebear	Ground truth: Grape Qwen-VL: gray gorilla DeepSeek-VL: square gorilla LLaVA: gorilla CogAgent: gorilla GPT-4V: gorilla GPT-4o: Kong	Ground truth: Blow off steam Qwen-VL: plugging away DeepSeek-VL: blowing in the wind LLaVA: blow off CogAgent: blow the snow away GPT-4V: cold shoulder GPT-4o: blow hot and cold	Ground truth: Ring a bell Qwen-VL: to ring a bell DeepSeek-VL: to ring the bell LLaVA: ring a bell CogAgent: a bell GPT-4V: sound the alarm GPT-4o: saved by the bell
Fine-grained image understanding of the emoji symbol	Emoji: 🌍 🐘	Emoji: 🐱 🏛️	Emoji: 😇 🛫 🙅 😁 👴 🤔 📝	Emoji: 📦 💸
	Ground truth: African elephant Qwen-VL: Earth elephant DeepSeek-VL: elephant LLaVA: elephant CogAgent: elephant on earth GPT-4V: worldwide GPT-4o: elephant	Ground truth: Caterpillar Qwen-VL: cat green pillar DeepSeek-VL: cat LLaVA: cat CogAgent: cat GPT-4V: catastrophe GPT-4o: cathedral	Ground truth: It is never too old to learn. Qwen-VL: fly by the seat of your pants DeepSeek-VL: grinning from ear to ear. LLaVA: Thinking helps a lot. CogAgent: Practice makes perfect. GPT-4V: an emotional rollercoaster GPT-4o: To err is human; to forgive, divine.	Ground truth: Receive a kickback Qwen-VL: think outside the box DeepSeek-VL:open the box and find money inside LLaVA: a box of money CogAgent: box outside GPT-4V: think outside the box GPT-4o: out of the box

Figure 15: Four typical problems the MLLMs suffer in English word and idiom tasks.

Problems	**Chinese idiom with four character**		**Chinese idiom with more than four characters**	
Couldn't identity ***homophonic characters***（Red color denotes the homophonic characters or sound-like characters）	Emoji: 🎃🐍🎃🍐	Emoji: 🌽🌨️♾️🏢	Emoji: 🎀🌩️🙅‍♀️🐓🤦‍♀️👂	Emoji: 🈴💧🙅‍♀️🍚#️⃣💧
	Ground truth: 难舍难离 Qwen-VL: 虎视蛇行 DeepSeek-VL:家徒四壁 LLaVA: 杯弓蛇影 CogAgent: 蛇蝎心肠 GPT-4V: 南瓜蛇果 GPT-4o: 南瓜蛇离	Ground truth: 玉宇琼楼 Qwen-VL:金碧辉煌 DeepSeek-VL: 高楼大厦 LLaVA:一举两得 CogAgent: 醉生梦死 GPT-4V:黄粱美梦 GPT-4o:五谷丰登	Ground truth: 捷雷不及掩耳 Qwen-VL:电闪雷鸣 DeepSeek-VL: 耳聪目明 LLaVA:拔鸡代猴 CogAgent:束手无策 GPT-4V:闻鸡起舞 GPT-4o:闻鸡起舞	Ground truth: 河水不犯井水 Qwen-VL:哭笑不得 DeepSeek-VL:一见钟情 LLaVA:泥菩萨过江 CogAgent:滴水穿石 GPT-4V:一波三折 GPT-4o:泪流满面
Suffer ***hallucination*** problem	Emoji: 🪑🥩🥚🐯	Emoji: 🤗🔥❌🆕	Emoji: 🙅‍♀️❓3️⃣7️⃣2️⃣🔟1️⃣	Emoji: 📿🏞️💃⚔️, IT📍🌊♂️
	Ground truth: 以肉啖虎 Qwen-VL:狗急跳墙 DeepSeek-VL:卧虎藏龙 LLaVA:虎口拔牙 CogAgent:狼心狗肺 GPT-4V:狐假虎威 GPT-4o:如坐针毡	Ground truth：抱火厝薪 Qwen-VL:喜新厌旧 DeepSeek-VL: 火光冲天 LLaVA:旧的不去新的不来 CogAgent:笑而不答 GPT-4V:新官上任三把火GPT-4o:笑里藏刀	Ground truth: 不问三七二十一 Qwen-VL: 三七二十一 DeepSeek-VL:问鼎中原 LLaVA:女大十八变 CogAgent: 三七二十一 GPT-4V:三心二意 GPT-4o:颠三倒四	Ground truth:项庄舞剑，意在沛公 Qwen-VL:背井离乡 DeepSeek-VL: 盲人摸象 LLaVA:投笔从戎 CogAgent:望洋兴叹 GPT-4V:珠光宝气 GPT-4o:草木皆兵，风声鹤唳
Multi-emoji to one character mapping	Emoji: 🥣🦩1️⃣0️⃣0️⃣0️⃣🧗	Emoji: 🌞🈳1️⃣0️⃣0️⃣0️⃣0️⃣🍐	Emoji: 🀄🈴🦁🥖, 1️⃣🥌1️⃣0️⃣0️⃣0️⃣🏅	Emoji: 🌅🌅🥶🍚🌃🌃⚪
	Ground truth:万壑千岩 Qwen-VL:一丝不苟 DeepSeek-VL:一声不吭 LLaVA:一鸣惊人 CogAgent:一见钟情 GPT-4V:一心一意 GPT-4o:差强人意	Ground truth: 晴空万里 Qwen-VL: 四大皆空 DeepSeek-VL: 艳阳高照 LLaVA: 空谷幽兰 CogAgent: 空空如也 GPT-4V:一暴十寒 GPT-4o:晴天霹雳	Ground truth: 中河失舟，一壶千金 Qwen-VL: 一举成名 DeepSeek-VL:画龙点睛 LLaVA:一言九鼎 CogAgent: 狡兔三窟 GPT-4V:朝三暮四 GPT-4o:狐假虎威	Ground truth: 朝朝寒食，夜夜元宵 Qwen-VL: 冰冻三尺非一日之寒 DeepSeek-VL: 有口难言 LLaVA:日出而作，日入而息 CogAgent:一日三秋 GPT-4V:破釜沉舟 GPT-4o:日月潭
Fine-grained image understanding of the emoji symbol	Emoji: 🐳🌈🆚🐦	Emoji: ⏬🦶🔼🤍	Emoji: 🤓🈚🙅‍♂️🗣️·🗣️🈚🙉	Emoji: 🍔🏷️🙅‍♂️🍔📓🚫😕🙎‍♀️?🚲💁‍♂️
	Ground truth: 精才绝艳 Qwen-VL: 鸟语花香 DeepSeek-VL: 掌上明珠 LLaVA: 海阔天空 CogAgent: 鲸吞蚕食 GPT-4V:东施效颦 GPT-4o:鲤鱼跃龙门	Ground truth: 抵足谈心 Qwen-VL: 心神不宁 DeepSeek-VL: 心悦诚服 LLaVA: 以退为进 CogAgent: 心服口服 GPT-4V: 下落不明 GPT-4o: 兵贵神速	Ground truth: 知无不言，言无不听 Qwen-VL: 无所不能 DeepSeek-VL: 笑口常开 LLaVA: 对牛弹琴 CogAgent: 对牛弹琴 GPT-4V:人山人海 GPT-4o:见不得人	Ground truth:止谤莫若自修 Qwen-VL: 十年寒窗 DeepSeek-VL:一波未平，一波又起 LLaVA: 一个巴掌拍不响 CogAgent: 一问三不知 GPT-4V:瓜田李下 GPT-4o:杯弓蛇影

Figure 16: Four typical problems the MLLMs suffer in Chinese idiom tasks.