Beyond Single Frames: Can LMMs Comprehend Temporal and Contextual Narratives in Image Sequences?

Anonymous ACL submission

Abstract

Large Multimodal Models (LMMs) have achieved remarkable success across various visual-language tasks. However, existing benchmarks predominantly focus on single-image understanding, leaving the analysis of image sequences largely unexplored. To address this limitation, we introduce StripCipher, a comprehensive benchmark designed to evaluate capabilities of LMMs to comprehend and reason over sequential images. StripCipher comprises a human-annotated dataset and three challenging subtasks: visual narrative comprehension, contextual frame prediction, and temporal narrative reordering. Our evaluation of 16 state-of-the-art LMMs, including GPT-4o and Qwen2.5VL, reveals a significant performance gap compared to human capabilities, particularly in tasks that require reordering shuffled sequential images. For instance, GPT-4o achieves only 23.93% accuracy in the reordering subtask, which is 56.07% lower than human performance. Further quantitative analysis discuss several factors, such as input format of images, affecting LMMs' performance in sequential understanding, underscoring the fundamental challenges that remain in the development of LMMs.

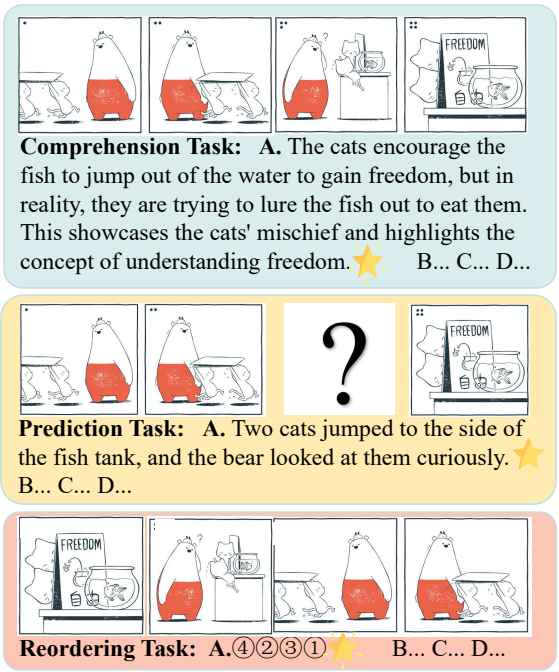

Figure 1: An example from the StripCipher dataset includes the correct answers for our three sub-tasks: comprehension, frame prediction, and reordering. All these tasks are presented as multiple-choice questions, with distractors excluded due to limited context. Star means correct answer.

1 Introduction

> *In the space between the panels, human imagination takes separate images and transforms them into a single idea.*
>
> — Scott McCloud (1993)

Recent advancements in Large Multimodal Models (LMMs), particularly GPT-4o (Hurst et al., 2024), have yielded significant breakthroughs in a various visual-language tasks, including image captioning (Liu et al., 2023b; Ghandi et al., 2024), visual question answering (Lu et al., 2023; Zhu et al., 2024), video understanding (Zhang et al., 2023; Maaz et al., 2024), and so on. LMMs have shown promising efficacy in processing and interpreting visual content (Yin et al., 2024), greatly enhancing their capacity to interact with the real world.

Despite recent advances, the capability of LMMs to process and reason over sequential images remains underexplored, even though sequential visual inputs are prevalent in real-world applications (Yang et al., 2024; Liu et al., 2024b). While existing benchmarks primarily evaluate LMMs on single images, often emphasizing surface-level understanding, they fail to address the complexities of sequential dependencies. In multi-image contexts, the ability to discern implicit meanings and

Benchmark	Task	Sequential	Reorder	#Frames	#Seq	#Cat
HCD (Radev et al., 2016)	Funniness Classification	No	No	50	0	1
MTSD (Cai et al., 2019)	Sarcasm Classification	No	No	9,638	0	3
HUB (Hessel et al., 2023)	Matching+Ranking+Explanation	No	No	651	0	1
MORE (Desai et al., 2022)	Sarcasm Explanation	No	No	3,510	0	1
DEEPEVAL (Yang et al., 2024)	Description+Title+Deep Semantics	No	No	1,001	0	6
AutoEval-Video (Chen et al., 2024a)	Open-Ended Question Answering	Yes	No	1,033	327	12
Mementos (Wang et al., 2024c)	Description Generation	Yes	No	8,124	699	1
STRIPCIPHER (Ours)	Prediction+Comprehension+Reorder	Yes	Yes	3,635	896	6

Table 1: Features and statistical information of STRIPCIPHER and prior related datasets. "Sequential" refers to whether image sequences are included. "Reorder" refers to whether benchmark tests the model's ability to understand images and reorder them. "#Frames" refers to the number of the frames(sub-image). "#Seq" refers to the number of image sequences. "#Cat" refers to the number of categories of the images.

contextual relationships is essential for comprehensive interpretation. In particular, understanding nuanced concepts such as sarcasm (Cai et al., 2019; Tang et al., 2024), humor (Patro et al., 2021; Hessel et al., 2023), and other multi-faceted deep meanings (Zhang et al., 2024a) requires reasoning beyond isolated frames. This gap highlights the need for a deeper investigation into the reasoning capabilities of LMMs over dynamic image sequences.

To address this gap, we introduce STRIPCIPHER, a novel benchmark designed to assess the reasoning ability of LMMs on temporal image sequences, including contextual structure, temporal relationships among images, and underlying semantics. STRIPCIPHER consists of three challenging subtasks: visual narrative comprehension, contextual frame prediction, and temporal frames reordering, as shown in Figure 1.

With STRIPCIPHER, we aim to advance the development of LMMs in temporal-visual comprehension while identifying their current limitations.

Our comprehensive evaluation of 16 state-of-the-art LMMs on STRIPCIPHER reveals a substantial performance gap between AI and human capabilities in sequential image comprehension, especially in the reordering task. Most notably, GPT-4o achieves only 23.93% accuracy in the reordering subtask and trails human performance by 30% in visual narrative comprehension. Further quantitative analysis identifies several key factors affecting the sequential understanding performance of LMMs, highlighting the fundamental challenges that remain in the development of LMMs.

2 Related Work

Large Multimodal Models Large language models (LLMs) have demonstrated strong performance in various natural language understanding and generation tasks (Dubey et al., 2024; Liu et al., 2024a; Ray, 2023). Building on the scaling law of LLMs, a generation of Large Multimodal Models (LMMs) has emerged, with LLMs serving as the backbone. These models (Team, 2025; Wang et al., 2024a; Zhu et al., 2024; Liu et al., 2023a; Wang et al., 2023; Ye et al., 2023) integrate visual features with language models using additional layers or specialized modules. Moreover, several closed-source LMMs (Reid et al., 2024; Driess et al., 2023; Yang et al., 2023b), including GPT-4o (Hurst et al., 2024), have demonstrated remarkable capabilities in handling complex multimodal inputs (Yue et al., 2024; Fu et al., 2023; Li et al., 2023). Beyond single-image LLMs, video LLMs (Zhang et al., 2025; Ye et al., 2024; Zhang et al., 2024c) further analyze and understand video content, which is essentially a continuous sequence of images. However, studies indicate that LMMs still encounter limitations in understanding implicit meanings (Yang et al., 2024; Liu et al., 2023c; Yang et al., 2023a), especially in the context of multiple sequential images, where research is lacking.

Visual Implicit Meanings Understanding Beyond studies on surface-level image understanding (Antol et al., 2015; Wang et al., 2022; Dong et al., 2022; Xia et al., 2023), recent works have shown that LMMs struggle implicit meaning understanding (Desai et al., 2022; Abu Farha et al., 2022; Hu et al., 2024). A recent study (Yang et al., 2024) further highlights a significant gap between AI and human comprehension of implicit meanings in images. However, these works are limited to single-image analysis. Multiple sequential images, arranged temporally, provide richer contextual information and serve as a bridge between static images and videos. Existing studies on sequential images have only focused on surface-level understanding (Chen et al., 2024a; Wang et al., 2024c).

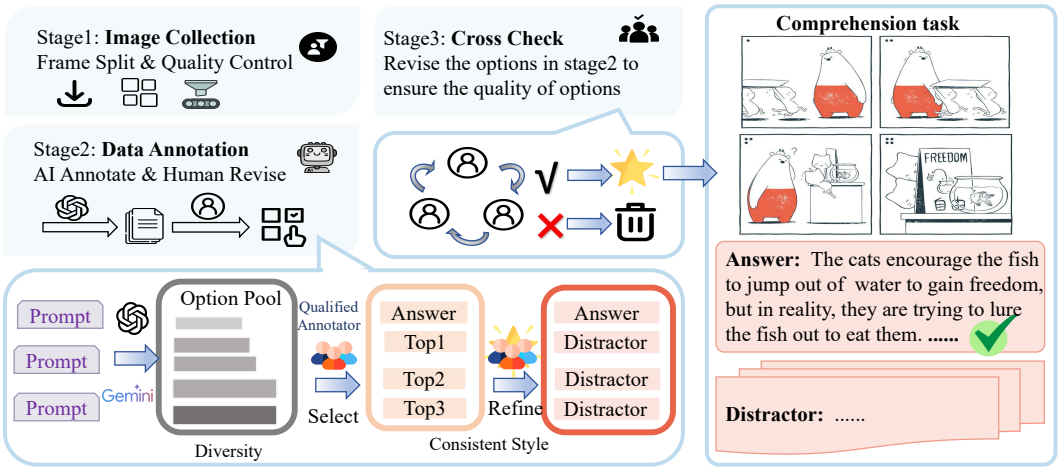

Figure 2: Schematic diagram of StripCipher dataset construction process including three stages: *Image Collection*, *Data Annotation* and *Cross Check*. Only comprehension task is displayed, the process of prediction task is the same as understanding. The reordering task only requires first-stage processing to ensure the ordering of frames is unique, and it does not require data annotation.

For instance, Wang et al. (2024c) collects image sequences and provides descriptions of their surface content and events without delving into deeper meanings. A detailed comparison with prior work is presented in Table 1, and the detailed description of categories and distributions covered by our method are illustrated in Appendix E.

3 Dataset and Task Overview

To investigate the capabilities of LMMs to comprehend sequential images, we introduce StripCipher, a novel benchmark consisting of three subtasks:

- **Visual Narrative Comprehension:** Examines whether models accurately interpret the narrative content of image sequences.
- **Contextual Frame Prediction:** Assesses the model reasoning ability to predict missing frames in image sequences based contextual.
- **Temporal narrative Reordering:** Evaluates whether models correctly infer and restore the chronological order of image sequences based causal temporal relationship.

The instructions for three subtasks are presented in Table 2. These subtasks provide a rigorous and multifaceted assessment of LMMs, offering insights into their strengths and limitations in sequential image understanding. In the following, we may use their full names or refer to them as *comprehension*, *prediction*, and *reordering* for simplicity.

Detailed statistics is displayed on Table 3. Overall, our proposed StripCipher includes 896 image sequences, with an average frame length of 4.09 of each sequence. The number of frames ranges from 3 to 8. Each task is designed in the form of multiple-choice questions, except for the reorder task, which also includes a question-answering format. Since its options are simple and well-defined, accuracy can be directly computed. In our tasks, the input format uniformly consists of images paired with textual prompts. Specifically, the comprehension task utilizes the whole images without split, the reordering task takes shuffled image sequence as input, and the frame prediction task involves masking second-to-last frame within

Comprehension: *What is happening in this comic strip? What is the implicit meaning? Based on your understanding, answer the question.*
Prediction: *Based on the overall story, what is happening in the blank frame (second-to-last)?*
Reordering: *The sequence of the comic strips provided below is incorrect. Your task is to find out the correct order of the comic strips based on the storyline and temporal logical relationship. Number each comic strip in the order they should appear, starting from 1.*

Table 2: Instruction for each subtasks.

Task	#Examples	Length	#Frames	#Type
Comprehension	680	30.53	2406	6
Prediction	600	19.07	2640	6
Reordering	890	4.09	3641	6

Table 3: Statistics of StripCipher. #Nums refers to the sum of samples. Length refers to the average of options. #Frames refers to the sum of frames. #Type refers to the sum of categories of images.

the image sequence. For the frame prediction task, we select the second-to-last frame, as it typically serves as a bridge between the preceding frames and the final frame. The start and end frames are generally more challenging to predict.

4 Dataset Construction

We construct our StripCipher dataset in a multi-step crowd-sourcing pipeline, including 1) annotator training, 2) data annotation, and 3) cross-check examination. An overall demonstration of our dataset construction pipeline is illustrated in Figure 2.

4.1 Image Source

We use silent comic strips, comic with panels and no dialogue, as our primary data source. As a distinct art form, comic strips often encapsulate complex narratives within concise visual sequences, addressing deeper themes such as social satire, humor, and inspiration. These characteristics make comic strips a particularly challenging medium for evaluating ability of LMMs to understand visual sequences. The dataset comprises samples from well-known comics, such as *Father and Son* and *Peanuts*, along with web-scraped images from `GoComics` [1], `Google`, and `Facebook` [2]. Initially, we collected $1,260$ images and then refined the dataset through a filtering process. We conducted a thorough manual inspection to eliminate unclear, toxic, overly simplistic images, along withmulti-panel comics lacking a clear temporal sequence. Moreover, comics with dialogue will also be removed to prevent the model from using OCR to understand the meaning of the comics through text rather than through images. As a result, the final dataset was reduced to 896 images.

4.2 Phase 1: Data Annotation.

Annotator Training We posted job descriptions on online forums and received over 50 applications from candidates with at least a Bachelor's degree. To ensure dataset quality, we provided training sessions that included online pre-annotation instructions and a qualification test to assess candidates' performance. Only those scoring above 95% were selected. candidates were assigned to one of two groups: annotators or inspectors. Ultimately, we hired 13 annotators and 7 inspectors for our data annotation process. To optimize efficiency and reduce costs, we implement a semi-automated pipeline for StripCipher annotation, leveraging `GPT-4o`[3]. Specifically, our data annotation process consists of two substeps: *answer creation* and *distractor generation*.

Answer Creation. The bottom panel of Figure 2 illustrates the process of our answer creation phase. Notably, only the comprehension task and frame prediction task need option annotation. The reordering task only necessitates using a program to randomly shuffle the frame order as the answer order, without the need for manual selection of the correct answer. We adopt an AI-assisted annotation approach in which human annotators refine pre-generated answers instead of creating them from scratch. Initially, we leverage `GPT-4o`, `GPT-4o-Mini` (Hurst et al., 2024) and Gemini (Reid et al., 2024) to generate diverse candidate answers. Human annotators then evaluate the image sequence and candidate answers, selecting the most appropriate ground truth. If none of the candidate answers is suitable, annotators are instructed to either refine a specific answer or create a new ground truth from scratch, which is about 28%.

Distractor Generation. We use candidates with plausible hallucinations from the previous sub-step as strong distractors. To ensure diversity in the multiple-choice options, we also prompt `GPT-4o`, `GPT-4o-mini` (Hurst et al., 2024) to generate intentionally incorrect responses as weak distractors. Typically, the responses of `GPT-4o-mini` are not very accurate and are mostly used as distractors. A detailed example of model annotation can be found in the appendix C Annotators are instructed to evaluate the quality of these distractors and select top-3 options, refining them if necessary. Finally, each ground truth is paired with three high-quality distractors for evaluation.

4.3 Phase 2: Cross-Check Examination

We implement a cross-check examination mechanism to ensure rigorous screening of high-quality annotations. During the data annotation process, hired inspectors review the annotated data and corresponding image sequences. If they encounter low-quality annotations, they have the option to reject them. Each annotation is reviewed by two

[1] https://www.gocomics.com/
[2] https://www.facebook.com/
[3] We use the `gpt-4o-2024-11-20` version for the data annotation process and subsequent evaluations in this work.

inspectors. If both inspectors reject the annotation, it is discarded, and the image is returned to the dataset for re-annotation. If an image sequence is rejected in two rounds of annotation, it suggests that this sample is not suitable for the current subtask (e.g., the meaning of the sample is unclear), and the image is subsequently removed from the subtask.

After annotation, both advanced annotators and inspectors, acting as final examiners, review the annotations to ensure they meet the required standards. Each annotation undergoes review by three examiners, who vote on whether to accept the annotated sample. Only the samples that receive a majority vote are approved. To ensure the quality of the examiners' work, we randomly sample 10% of the annotations for verification.

4.4 Data Composition

It is important to note that the reordering subtask does not require human annotation, as described in the previous process. For this subtask, we select suitable image sequences based on the criterion that the correct ordering must be unique. To ensure this, we conduct a manual review to verify that each sequence follows a logically unambiguous order. A script is then run to perform the initial splitting of panels within specific comics, followed by a random shuffling of these panels. Human annotators are tasked with verifying the format and quality of the frames to ensure they meet the required standards. These processed image sequences serve as the evaluation data for the reordering subtask.

The final version of our 32-day annotated STRIPCIPHER contains 896 items (see Table 2), encompassing three subtasks: *visual narrative comprehension*, *contextual frame prediction*, and *temporal narrative reordering*. In each of these subtasks, each sample consists of an image sequence paired with a multiple-choice question offering four options. The evaluated LMMs are required to select the option they deem most appropriate from the four. More information and examples of STRIPCIPHER can be found in Appendix D.

5 Experiments

5.1 Models

To comprehensively evaluate on LMMs, we conducted zero-shot inference across both commercial and open-source models. Our evaluation suite includes leading commercial models GPT-4o (Hurst et al., 2024) and Gemini1.5-Pro (Anil et al., 2023) alongside state-of-the-art open-source alternatives of varying scales: Qwen2.5-VL (Team, 2025), Qwen2-VL (Wang et al., 2024b), LLaVA-v1.6 (Liu et al., 2023b), CogVLM (Wang et al., 2023), MiniCPM-o-2.6 (Yao et al., 2024), mPlug-Owl2 (Ye et al., 2023), InternVL2v5 (Chen et al., 2024b),LLaVA-NEXT-Video (Zhang et al., 2024b) and Cambrian (Tong et al., 2024). Besides, Janus-Pro (Chen et al., 2025), which unifies multimodal understanding and generation, is included to test the abilities between Unified Model and Vision Language Model. This diverse selection enables us to analyze how model scale, architecture, and training approaches influence comic comprehensive capabilities.

5.2 Experimental Details

The task prompts is displayed in Table 2. For visual narrative comprehension task, model is provided with the whole image. But for next-frame prediction and multi-frame sequence reordering task, LMMs infer with image sequences. The hyper-parameters for each LMMs in the experiments including possible settings are detailed in Appendix B. Furthermore, to assess human capabilities in these tasks, we randomly select 100 questions from the dataset for each task and instruct human evaluators to answer. This allows us to benchmark the performance of human participants against our models, offering a thorough comparison of both human and LMMs proficiency in these specific tasks.

5.3 Main Results

Our comprehensive evaluation reveals that while LMMs show promising capabilities in comprehension and prediction tasks, they significantly underperformed in sequence reordering tasks. Moreover, there remains a substantial performance gap between current models and human performance across all tasks. Unified Model underperformed than Vision Language Model.

Contextual Frame Prediction The frame prediction task appears to be the most tractable among the three tasks. GPT-4o achieves the highest score of 69.95%, followed closely by Qwen2-VL at 64.00%.This demonstrates that the performance gap between closed and open-source models is relatively small for this task. However, Janus-Pro perform notably below expectations (27.50%), pos-

Models	Backbone	#Params	I-Video	Comprehension	Prediction	Reordering		AVG
TYPE						choice	VS generation	
Human				80.00	82.00	86.00	80.00	82.00
Closed-Model								
GPT-4o-mini (Hurst et al., 2024)	-	-		53.23	56.33	**26.07**	8.45	36.02
Gemini1.5-Pro (Reid et al., 2024)	-	-		49.56	67.83	25.51	**32.02**	43.73
GPT-4o (Hurst et al., 2024)	-	-		**61.60**	**69.95**	25.17	23.93	**45.16**
Open-Source								
Janus-Pro (Chen et al., 2025)	DeepSeek-LLM-7b-base	7B	No	27.50	27.50	26.07	*	20.27
mPlug-Owl2 (Ye et al., 2023)	LLaMA2	8B	No	30.74	31.17	25.06	0.56	21.88
LLaVA-v1.6 (Liu et al., 2023b)	Vicuna-v1.5	7B	No	34.41	43.50	26.29	3.37	26.89
LLaVA-NeXT-Video (Zhang et al., 2024b)	Vicuna-v1.5	7B	Yes	45.74	44.50	23.71	*	28.49
CogVLM (Wang et al., 2023)	Vicuna-v1.5	17B	No	34.26	56.00	24.83	*	28.77
LLaVA-v1.6 (Liu et al., 2023b)	Vicuna-v1.5	13B	No	46.03	46.50	27.98	2.58	30.77
LLaVA-v1.6 (Liu et al., 2023b)	Vicuna-v1.5	34B	No	52.94	50.83	25.62	2.13	32.88
Cambrian (Tong et al., 2024)	Vicuna-v1.5	13B	No	45.59	55.00	26.85	4.94	33.10
Qwen2.5VL (Team, 2025)	Qwen2.5	3B	Yes	50.59	58.83	27.75	1.91	34.77
InternVL2v5 (Chen et al., 2024b)	Intern	26B	Yes	60.92	65.17	24.61	2.58	38.32
MiniCPM-o 2.6 (Yao et al., 2024)	Qwen2.5	7B	Yes	56.18	**65.83**	26.85	5.51	38.59
Qwen2.5VL (Team, 2025)	Qwen2.5	7B	Yes	56.03	64.00	29.21	**11.01**	40.06
Qwen2VL (Wang et al., 2024a)	Qwen2	7B	Yes	**58.53**	63.00	**31.91**	9.44	**40.72**

Table 4: Model performance comparison across different architectures and scales. The table is sorted by accuracy. I-Video indicates whether the large multimodal models (LMMs) support video input. Scores are reported as percentages (%). Prediction, Comprehension, and Reordering correspond to visual narrative understanding, next-frame prediction, and multi-frame reordering, respectively. The * denotes that the model failed on the corresponding task. AVG represents the average score across the four scores including the failed one. Bolded values indicate the highest scores among closed-source and open-source models.

sibly due to its unified model architectural.

Visual Narrative Comprehension For visual narrative comprehension, we observe a similar pattern but with generally lower scores. GPT-4o leads with 61.60%, while other models show varying degrees of capability.

Temporal Narrative Reordering The frame reordering task proves to be the most challenging, with all models performing significantly below human capability. Even the best-performing models struggle to exceed 30% accuracy, with many achieving scores around $25-26\%$, which is slightly higher than random selection. Notably, several models (marked with *) are unable to perform this task due to their architectural limitations in processing multiple images simultaneously. For these models, we attempted to accommodate their single-image constraint by concatenating multiple frames horizontally into a single image, with white margins serving as frame boundaries. However, this workaround appears to be suboptimal, as these models likely struggle to properly distinguish individual frame boundaries and maintain the semantic independence of each frame, ultimately leading to their poor performance on the reordering task.

The poor performance on reordering task suggests that current LMMs, regardless of their scale or architecture, have not yet developed robust capabilities for understanding temporal relationships and sequential logic in visual narratives.

6 Analysis

Our analysis addresses the following questions:

Does fine-tuning with reorder task help? Yes, it does. We fine-tune Qwen2-VL using 3,160 samples for one epoch. This not only significantly improves performance on the reordering VQA task but also enhances comprehension tasks. To construct the training dataset, we applied data augmentation to 790 images using the reorder task. Specifically, we randomly shuffled the sequence of images four times, generating a total of approximately 3,160 distinct samples. For evaluation on the reorder task, we used only the remaining 100 samples. For the comprehension task, we conducted a full test set evaluation, as the training data provided only images without any analytical content. Meanwhile, we excluded the frame prediction task from testing due to potential data leakage. The experimental results are presented in the table 5. Overall, our reordering data is useful for fine-tuning, as it can enhance the LMMs to reason on sequential images. However, the ultimate performance still depends on the base capability of the model. From the table, it can be seen that Qwen2.5-VL benefits greatly from fine-tuning, with improvements not only in the reordering task but also in comprehen-

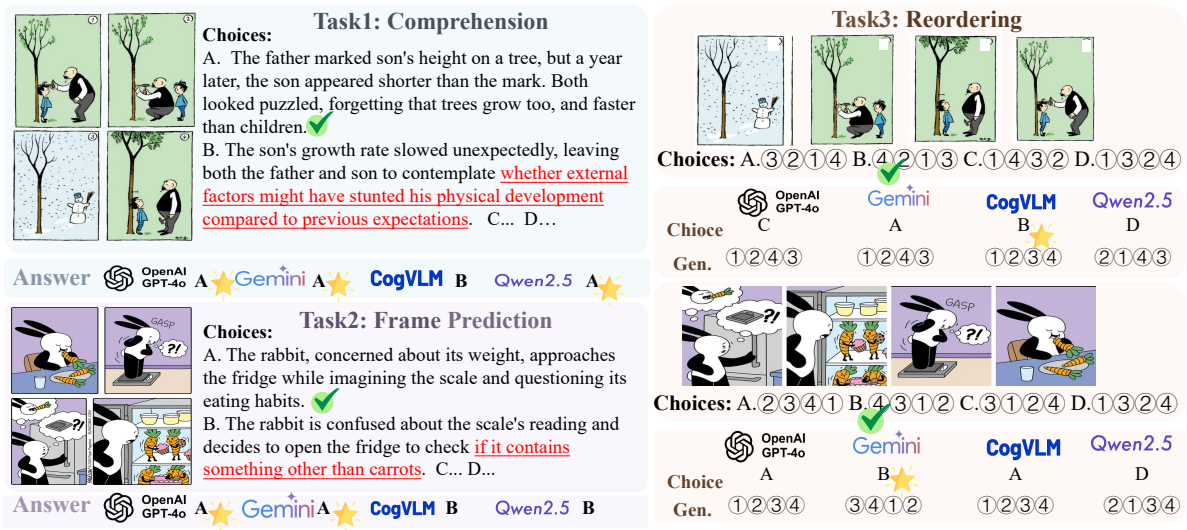

Figure 3: Sample outputs of our three tasks generated by different vision language models, along with gold truth. We highlight errors in distractors.

sion tasks. The improvement in generation tasks in the reordering task is much larger than that in choice tasks. We believe this is because the generation task's instructions are very challenging, and the model has not been trained on them before, leading to difficulty in solving them. On the other hand, for choice tasks, the model can make educated guesses based on the options provided. Due to the relatively small amount of training samples, the model's improvement in choice tasks is limited. In contrast, LLaVA shows improvement only in the reorder-choice task after training.

Tasks	Comprehension	Reordering-G	Reordering-C
Qwen2-VL	58.53	6.00	31.00
+finetune	62.94	31.00	38.00
LLaVA-1.6	34.41	3.00	26.00
+finetune	33.82	2.00	32.00

Table 5: Performance on Qwen2-VL-7B and LLaVA-1.6-7B finetuned with reordering task data. Reordering-C refers to the Reordering-choice. Reordering-G refers to the Reordering-generation

Does GPT-4o understand sequence images as well as humans? While GPT-4o achieves the highest performance among all tested models, there remains a substantial gap between its capabilities and human performance, particularly in novel tasks like frame reordering. Through our preliminary data annotation experiments, we observed that while GPT-4o can comprehend basic comic content and provide interpretations, it frequently generates hallucinated content and struggles with comics that depict unconventional or imaginative scenarios rarely encountered in real life.

In the visual narrative comprehension and next-frame prediction tasks, the multiple-choice format allows models to leverage similarity matching between options. Our investigation revealed that model performance is heavily influenced by the quality of distractor options. In initial experiments with weak distractors (generated using GPT-4o with instructions to provide distractors with hallucinations, the prompt is followed HalluEval), the model achieved accuracy rates up to 90%. Upon analysis, we found these initial distractors were too obviously incorrect or irrelevant to the comic content, making the selection task trivial. To address this limitation, we carefully curated a new set of challenging distractors. With these enhanced distractors, performance of GPT-4o decreased significantly to more realistic levels (61.60% for understanding and 69.95% for prediction), better reflecting the true challenges in comic comprehension. The scores obtained from multiple-choice questions with semantically transparent options tend to be inflated. In subsequent reordering tasks, where options lack explicit semantic meanings, coupled with open-ended questions, the scores provided a more authentic assessment of the LMMs.

Does input format of images influence performance? Considering the distinct computational pathways that LMMs employ in processing individual versus multiple images, we designed following experiments to measure the differential impact of varied input formats using Qwen2.5VL as our test case. We compared three input formats: (1)

whole image - the entire comic strip as a single image, (2) sequential frames - individually separated frames input in order, and (3) shuffled sequence - separated frames input in random order. Table 6 shows surprisingly consistent performance across all three formats (56.03%, 54.56%, and 57.65% respectively).

This consistency suggests that while separated frames might theoretically help models extract clearer information from each panel and avoid visual confusion from complex layouts, the current video-like processing mechanism used by LMMs for multiple images might not fully capitalize on these advantages. The similar performance with shuffled sequences further indicates that models might rely more on individual frame content rather than sequential relationships for understanding tasks.

Input	GPT-4o-mini	Qwen2.5-VL
Whole Image	53.23	56.03
Image Sequence	51.03	54.56
Shuffled Sequence	49.56	57.65

Table 6: Performance with different input format for understanding task.

Does implicit meaning help reordering task? To investigate whether poor reordering performance stems from inadequate semantic understanding, we enhanced the reordering task by providing explicit semantic annotations along with shuffled images. As shown in Table 7, this additional semantic information only marginally improved performance (from 30.01% to 32.54%). This modest improvement suggests that the bottleneck in reordering tasks lies not in semantic understanding but in the fundamental capability to reason about temporal and logical sequences in visual narratives.

Input	GPT-4o-mini	Qwen2.5-VL
Shuffled image	26.07	30.01
+Meaning	28.40	32.54

Table 7: Performance with enhanced data (correct answer for comprehension task) for reordering task.

How does model size affect? In this section, we will discuss the relationship between model parameters size and reasoning performance for tasks due to the scaling law. We examine on LLaVA and Qwen2.5VL, from 3B scale to 34B scale. There is a clear scaling effect across model sizes, as demonstrated by LLaVA1.5's performance improving from 34.41% (7B) to 46.03% (13B) to 52.94% (34B). This suggests that model scale plays a crucial role in comprehending implicit meanings in visual narratives. Figure 4 provide a visual representation of the performance trend for five models.

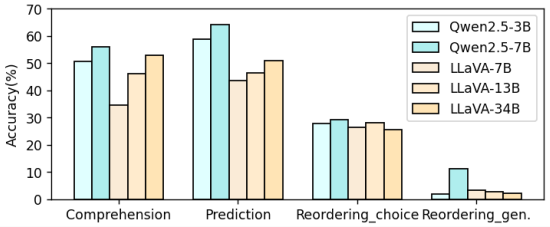

Figure 4: Comparison of the accuracy results between Qwen2.5-3B vs Qwen2.5-7B and LLaVA-1.6-7B vs LLaVA-1.6-13B vs LLaVA-1.6-34B

Where do LMMs fail? We present sample outputs of three tasks generated by vision language models (VLMs) in Figure 3. These images are easy for human but hard for VLMs. VLMs can understand one comic strip but they can still make mistakes with reordering task.

7 License and Copyright.

We used original web links to comic images without infringing on their copyright. This work is licensed under a CC BY-NC license. We will open-source all related code for processing image sequences and frames to facilitate the reproducibility of our evaluated image sequences. All annotators participated voluntarily in the annotation process and were provided fair compensation.

8 Conclusion

We present StripCipher, a comprehensive benchmark for evaluating Large Multimodal Models' capabilities in visual comic sequence reasoning. Our benchmark comprises meticulously curated and human-AI annotated tasks spanning visual narrative comprehension, next-frame prediction, and multi-frame sequence reordering. Through extensive evaluations of state-of-the-art LMMs, we identify significant performance gaps between AI systems and human capabilities in comic strip understanding. These findings underscore the considerable challenges that remain in developing AI systems capable of deep visual semantic understanding comparable to human cognition.

Limitations

Limited Availability of Comic Strips: Our dataset contains a relatively small number of samples due to the scarcity of standalone short-story comic strips available online. Most comics are either serialized narratives or dialog-driven, making it challenging to collect a diverse set of independent stories.

Narrow Focus on Comics: While our dataset captures the narrative structure of comic strips, it does not encompass the full spectrum of sequential visual storytelling, such as photographic sequences, instructional diagrams, or movie storyboards. Future work could extend beyond comics to explore a broader range of visual sequences.

Limited Training Data for Fine-Tuning: Our findings indicate that fine-tuning significantly enhances model performance on the reordering task. However, the limited availability of training data constrains the model's ability to fully develop temporal reasoning skills. Expanding the dataset or incorporating alternative sources, such as video sequences, could further improve performance.

Ethics Statement

The datasets used in our experiment are publicly released and labeled through interaction with humans in English. In this process, user privacy is protected, and no personal information is contained in the dataset. The scientific artifacts that we used are available for research with permissive licenses. And the use of these artifacts in this paper is consistent with their intended use. Therefore, we believe that our research work meets the ethics of ACL.

References

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. 2022. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 802–814, Seattle, United States. Association for Computational Linguistics.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling.

Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024a. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, pages 179–195. Springer.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence (Volume 1: Long Papers)*, pages 10563—-10571. The Symposium on Educational Advances in Artificial Intelligence.

Qingxiu Dong, Ziwei Qin, Heming Xia, Tian Feng, Shoujie Tong, Haoran Meng, Lin Xu, Zhongyu Wei, Weidong Zhan, Baobao Chang, Sujian Li, Tianyu Liu, and Zhifang Sui. 2022. Premise-based multimodal reasoning: Conditional inference on joint textual and visual clues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 932–946, Dublin, Ireland. Association for Computational Linguistics.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al.

2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2024. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3):62:1–62:39.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Zhe Hu, Tuo Liang, Jing Li, Yiren Lu, Yunlai Zhou, Yiran Qiao, Jing Ma, and Yu Yin. 2024. Cracking the code of juxtaposition: Can ai models understand the humorous contradictions. *ArXiv*, abs/2405.19088.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. 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*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Ziqiang Liu, Feiteng Fang, Xi Feng, Xeron Du, Chenhao Zhang, Noah Wang, yuelin bai, Qixuan Zhao, Liyang Fan, CHENGGUANG GAN, Hongquan Lin, Jiaming Li, Yuansheng Ni, Haihong Wu, Yaswanth Narsupalli, Zhigang Zheng, Chengming Li, Xiping Hu, Ruifeng Xu, Xiaojun Chen, Min Yang, Jiaheng Liu, Ruibo Liu, Wenhao Huang, Ge Zhang, and Shiwen Ni. 2024b. II-bench: An image implication understanding benchmark for multimodal large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. 2023. The multimodal fusion in visual question answering: a review of attention mechanisms. *PeerJ Comput. Sci.*, 9:e1400.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Badri N. Patro, Mayank Lunayach, Deepankar Srivastava, Sarvesh, Hunar Singh, and Vinay P. Namboodiri. 2021. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 576–585. IEEE.

Dragomir Radev, Amanda Stent, Joel Tetreault, Aasish Pappu, Aikaterini Iliakopoulou, Agustin Chanfreau, Paloma de Juan, Jordi Vallmitjana, Alejandro Jaimes, Rahul Jha, and Robert Mankoff. 2016. Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 475–479, Portorož, Slovenia. European Language Resources Association (ELRA).

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with vi-

sual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742, Mexico City, Mexico. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5-vl.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Ruonan Wang, Yuxi Qian, Fangxiang Feng, Xiaojie Wang, and Huixing Jiang. 2022. Co-vqa: Answering by interactive sub question sequence. *arXiv preprint arXiv:2204.00879*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. 2024c. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*.

Heming Xia, Qingxiu Dong, Lei Li, Jingjing Xu, Tianyu Liu, Ziwei Qin, and Zhifang Sui. 2023. ImageNetVC: Zero- and few-shot visual commonsense evaluation on 1000 ImageNet categories. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2009–2026, Singapore. Association for Computational Linguistics.

Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023a. Mm-bigbench: Evaluating multimodal models on multimodal content comprehension tasks. *arXiv preprint arXiv:2310.09036*.

Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can large multimodal models uncover deep semantics behind images? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1898–1912, Bangkok, Thailand. Association for Computational Linguistics.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023b. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

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

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *Preprint*, arXiv:2311.04257.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xinrun Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, Ge Zhang, and Shiwen Ni. 2024a. Can mllms understand the deep implication behind chinese images? *CoRR*, abs/2410.13854.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*,

pages 543–553. Association for Computational Linguistics.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Appendix

A Model

Our evaluation suite includes leading commercial models GPT-4o (Hurst et al., 2024) and Gemini1.5-Pro (Anil et al., 2023) alongside state-of-the-art open-source alternatives of varying scales: Qwen2.5-VL (Team, 2025), Qwen2-VL (Wang et al., 2024b), LLaVA-v1.6 (Liu et al., 2023b), CogVLM (Wang et al., 2023), MiniCPM-o 2.6 (Yao et al., 2024), mPlug-Owl2 (Ye et al., 2023), InternVL2v5 (Chen et al., 2024b),LLaVA-NEXT-Video (Zhang et al., 2024b) and Cambrian (Tong et al., 2024). Besides, Janus-Pro (Chen et al., 2025), which unifies multimodal understanding and generation, is included to test the abilities between Unified Model and Vision Language Model.

B Model Hyper-parameter Details

We use the default hyper-parameter values of the models. In the LLaVa-1.5-7B and LLaVa-1.5-13B, the temperature is set to 0.2. For MiniGPT-4, the temperature is set to 1.0, and num_beams is also set to 1.0. The temperature for mPlug-Owl-2 is set to 0.7. For CogVLM, the temperature is set to 0.4, top_p is set to 0.8, and top_k is set to 1.0.

In the LLaVa-1.6-7B, LLaVa-1.6-13B and LLaVa-1.6-34B, the temperature is set to 0.2. In the Qwen2-VL-7B, Qwen2.5-VL-3B and Qwen2.5-VL-7B, the temperature is set to 0.01, top_p is set to 0.001, and top_k is set to 1. For CogVLM-17B, the temperature is set to 0.4, top_p is set to 0.8, and top_k is set to 1.0. For InternVL2-26B, do_sample is set to False. For Cambrian-13B, the temperature is set to 0.2.

C Annotation

The following listed prompts are used to construct data. By instructing to different LMMs, we can obtain option candidate pool.

Prompt1: *"""What happened in comic strip? Conclude the whole story, then carefully analyze the implicit meaning of comic. Ouput in 35 words."""*

Prompt2: *""" You are now a mature hallucination generator. Please generate one strong distractor option for the following question. You can use any method you have learned that is misleading for the given question."""*

Prompt3: *""" Task Overview: Strive to understand this story and analyze its implicit meaning, then complete multi-choice question for test. You should act in two roles to complete tasks. Image Context: Pic1: The first picture shows the complete comic strip. Read it from left to right, top to bottom, to understand the full narrative arc. Pic2: The second picture is the second-to-last frame from Pic1, which is the target frame in Task 1.*

###Role 1 - Excellent Comic Analysis Expert:

Task 1: Contextual Scene Description Question 1: Based on the overall story, what is happening in the second-to-last frame (Pic2) of the comic strip? Requirements: 1. Provide a clear and detailed description of the key visual elements, characters, relationships and actions in Pic2. 2. Ensure narrative continuity with the events of the entire comic. 3. Output this as the right option for Task 2 with 30-40 words.

Task 2: Implicit Meaning Analysis Question 2: What happened in comic strip (Pic1)? Describe the whole story in detail, then analyze its implicit meaning. Requirements: 1. Describe the whole story in detail and analyze its implicit meaning and sentiment. 2. Provide three sentences with 40-50 words as right option for Task 4.

###Role 2 - Strong Distractor Options Generator: Task 3: Frame Scene Options Generation Question 1: Based on the overall story, what is happening in the second-to-last frame (Pic2) of the comic strip? Requirements: 1. Generate three plausible but incorrect options for #Question 1#. 2.The length of each option should be similar with the correct answer from Task 1 (around 30-40 words)! 3. Ensure that the incorrect options are consistent with the overall story but misinterpret the events of Pic2.

Task 4: Comic Strip Analysis Options Generation #Question 2:# What happened in comic strip (Pic1)? Describe the whole story in detail, then analyze its implicit meaning. Requirements: 1. Generate three plausible but incorrect options for #Question 2#. 2. Each incorrect option should be composed of 3 sentences and share the same length with the correct answer from Task 2! 3. Avoid obviously wrong or nonsensical answers.

#Please strictly adhere to the word count requirement! The final output should be in the following format:# #Question 1:# Based on the overall story, what is happening in the second-to-last frame (Pic2) of the comic strip? #Reasoning Chain 1:# [Describe the story in sequence.]

#Options in Q1:# A. [Hallucination option with 30-40 words from Task 3] ### B. [Hallucination

option with 30-40 words from Task 3] ### C. [Hallucination option with 30-40 words from Task 3] ### D. [Right option with 30-40 words from Task 1] ###

#Right Answer 1:# D

#Question 2:# What happened in comic strip (Pic1)? Describe the whole story in detail. Carefully analyze the implicit meaning of comic.

#Reasoning Chain 2:# [Reason step by step here.]

#Options in Q2:# A. [Hallucination option with 40-50 words from Task 4] ### B. [Hallucination option with 40-50 words from Task 4] ### C. [Hallucination option with 40-50 words from Task 4] ### D. [Right option with 40-50 words from Task 2] ###

#Right Answer 2:# D """

prompt4: *"""Predict what happened in the blank panel? Output in 35 words."""*

prompt5: *"""You are now a mature hallucination generator. Please generate one strong distractor option for the following question. You can use any method you have learned that is misleading for the given question. Question: Predict what happened in the blank panel. Please output with 35 words without any additional text or explanation."""*

D Examples

Here is an example of data construction in Figure 7.

E Category

We evaluated the performance of LLMs on comprehension tasks across 6 categories of comic strip.

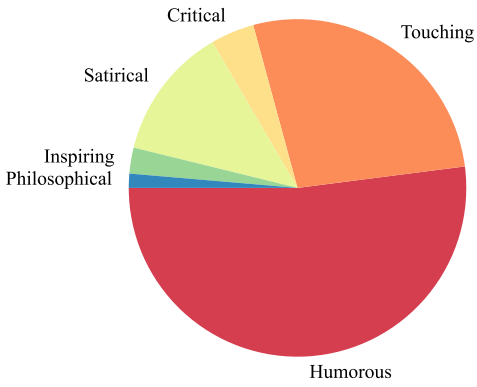

Figure 5: The distribution of six categories of StripCipher.

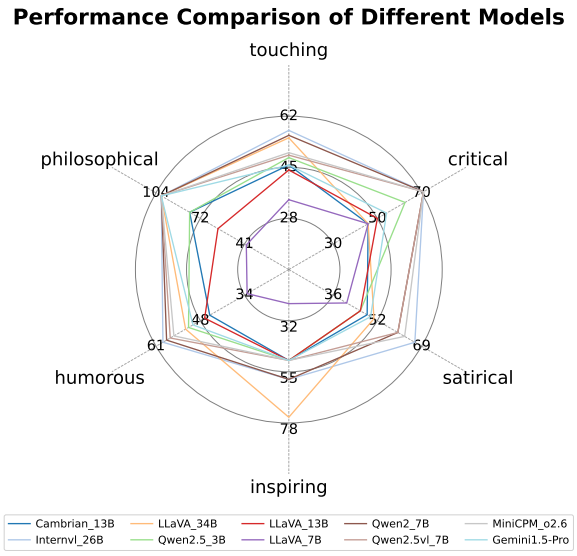

Figure 6: Comprehension task performance comparision of different LLMs on different categories.

Category	Definition
Satirical	The comic uses irony, exaggeration, or ridicule to criticize social, political, or cultural issues. It often highlights contradictions, hypocrisy, or absurdity in a way that provokes thought or debate. The humor may be sharp or biting but serves a critical purpose.
Inspiring	The comic presents a positive or uplifting message, often encouraging personal growth, motivation, or perseverance. It may depict acts of kindness, success against adversity, or wisdom that encourages the reader to strive for betterment.
Touching	The comic evokes emotions such as empathy, nostalgia, or affection. It may explore themes of love, friendship, loss, or family bonds, aiming to create a sentimental or heartfelt response from the audience.
Philosophical	The comic explores deep, abstract, or existential ideas about life, morality, meaning, or human nature. It prompts the reader to reflect on profound questions, often using metaphors or thought-provoking dialogue rather than direct humor or emotion.
Critical	The comic highlights flaws or problems in society, institutions, or human behavior with a serious or analytical tone. Unlike satire, which uses humor as a tool for critique, a critical comic may adopt a more straightforward, serious, or thought-provoking approach to expose issues and encourage awareness.
Humorous	The comic's primary goal is to entertain and amuse the audience. It relies on lighthearted jokes, wordplay, or visual gags without necessarily conveying a deeper message or critique. The tone is playful, aiming for laughter rather than serious reflection.

Table 8: The types and definition of the categories in StripCipher.

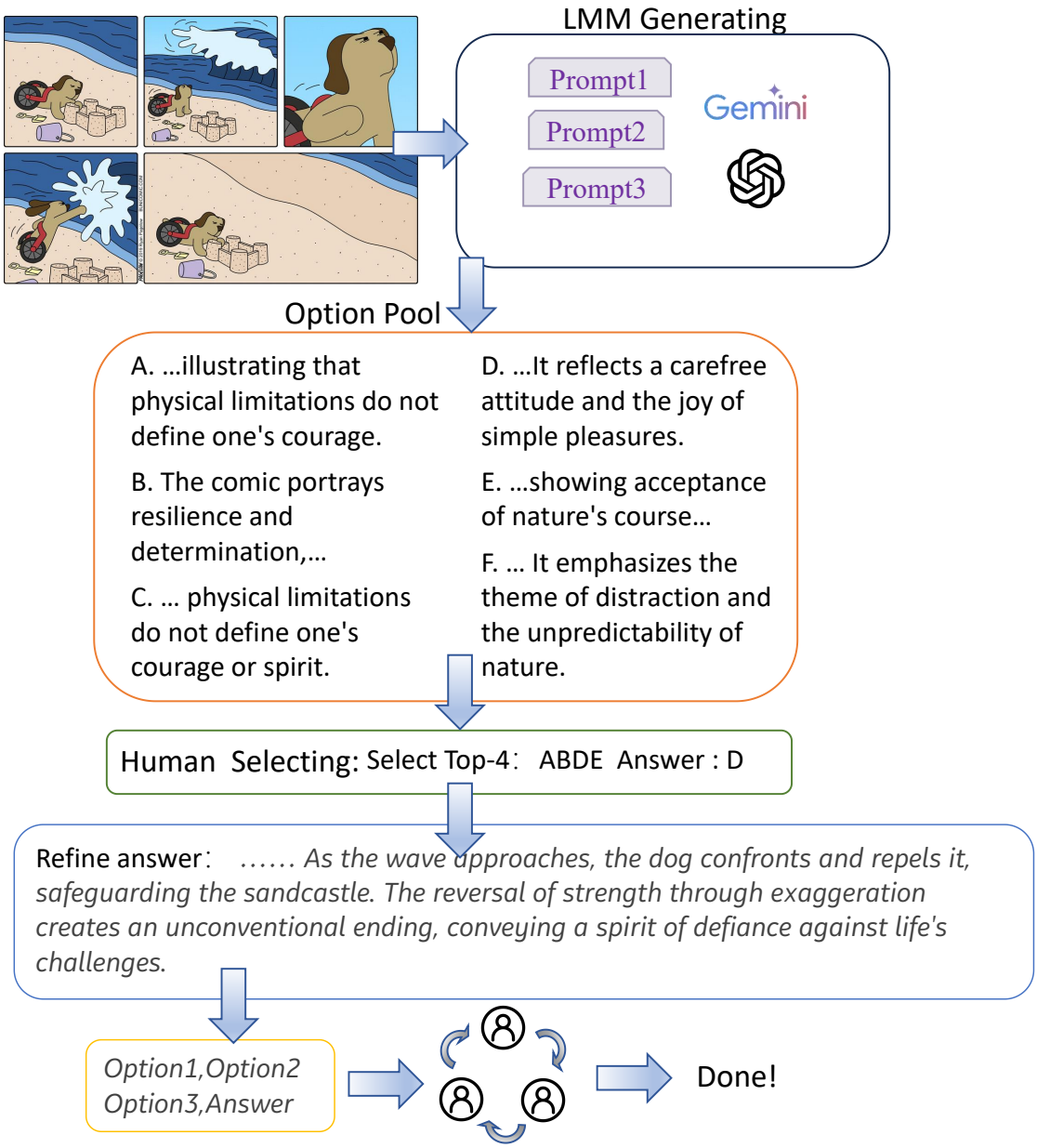

Figure 7: An detailed example of data construction.