Exploring the Capability of Multimodal LLMs with Yonkoma Manga: The YManga dataset and Its Challenging Tasks

Qi Yang*, Jingjie Zeng*, Liang Yang†, Zhihao Yang, Hongfei Lin
School of Computer Science and Technology,
Key Laboratory of Social Computing and Cognitive Intelligence,
Dalian University of Technology, China
{qiyang, jjtail}@mail.dlut.edu.cn, {liang, yangzh, hflin}@dlut.edu.cn

Abstract

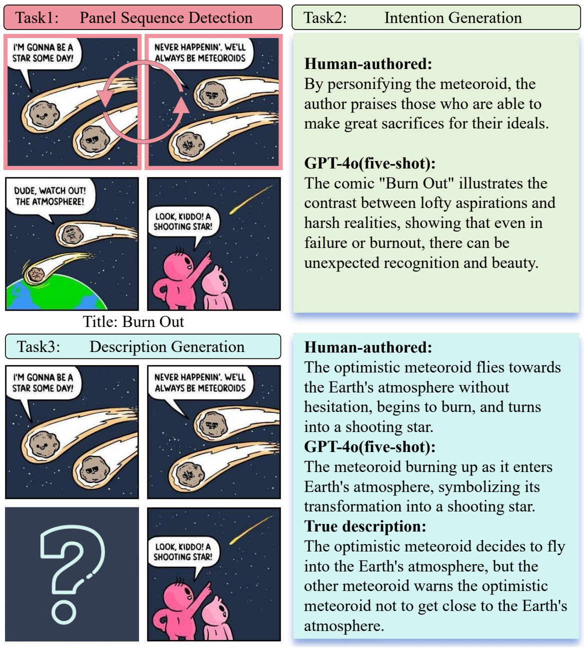
Yonkoma Manga, characterized by its four-panel structure, presents unique challenges due to its rich contextual information and strong sequential features. To address the limitations of current multimodal large language models (MLLMs)[1] in understanding this type of data, we create a novel dataset named YManga from the Internet. After filtering out low-quality content, we collect a dataset of 1,015 yonkoma strips, containing 10,150 human annotations. We then define three challenging tasks for this dataset: panel sequence detection, generation of the author's creative intention, and description generation for masked panels. These tasks progressively introduce the complexity of understanding and utilizing such image-text data. To the best of our knowledge, YManga is the first dataset specifically designed for yonkoma manga strips understanding. Extensive experiments conducted on this dataset reveal significant challenges faced by current multimodal large language models. Our results show a substantial performance gap between models and humans across all three tasks.[2]

1 Introduction

Yonkoma Manga, it will be referred to as yonkoma below for the sake of simplicity, also known as 4-koma Manga, which is a very regular comic data, originated in Japan and consists of four panels of equal size. Each panel typically follows this structure:

(a) Introduction: Sets up the scene or premise.
(b) Development: Builds on the initial setup.
(c) Twist: Introduces an unexpected turn.
(d) Conclusion: Delivers the resolution.

This format provides a concise and often humorous narrative within a limited space.


Task1: Panel Sequence Detection
I'M GONNA BE A STAR SOME DAY!
NEVER HAPPENIN'. WE'LL ALWAYS BE METEOROIDS
DUDE, WATCH OUT! THE ATMOSPHERE!
LOOK, KIDDO! A SHOOTING STAR!
Title: Burn Out
Task2: Intention Generation
Task3: Description Generation
I'M GONNA BE A STAR SOME DAY!
NEVER HAPPENIN'. WE'LL ALWAYS BE METEOROIDS
LOOK, KIDDO! A SHOOTING STAR!

Human-authored:
By personifying the meteoroid, the author praises those who are able to make great sacrifices for their ideals.

GPT-4o(five-shot):
The comic "Burn Out" illustrates the contrast between lofty aspirations and harsh realities, showing that even in failure or burnout, there can be unexpected recognition and beauty.

Human-authored:
The optimistic meteoroid flies towards the Earth's atmosphere without hesitation, begins to burn, and turns into a shooting star.
GPT-4o(five-shot):
The meteoroid burning up as it enters Earth's atmosphere, symbolizing its transformation into a shooting star.
True description:
The optimistic meteoroid decides to fly into the Earth's atmosphere, but the other meteoroid warns the optimistic meteoroid not to get close to the Earth's atmosphere.

Figure 1: We defined three tasks on YManga: (1) Panel Sequence Detect (PSD); (2) Author's Intention Generation (IG); (3) Description Generation (DG). Additionally, the last two tasks are further divided based on the presence or absence of manually annotated panel descriptions.

Yonkoma data, as showing in figure 1, exhibits unique structures and modes of information expression. Existing large-scale image-text datasets, such as MSCOCO (Lin et al., 2014), ImageNet (Deng et al., 2009), typically pair images with text in a straightforward manner, lacking sufficient contextual information. This simplistic correspondence limits the expressive capacity of these datasets in complex scenarios. In contrast, yonkoma data provides stronger contextual relevance and coherence, not only containing textual elements like titles but also delivering rich narrative information through sequential panels.

Furthermore, yonkoma data features a much higher information density compared to video data.

*Equal contribution.
†Corresponding author.
[1] Specifically refers to LLMs that can process image and text information.
[2] https://github.com/yangqi1725/YManga

Findings of the Association for Computational Linguistics: EACL 2024, pages 8672–8687
November 12-16, 2024 ©2024 Association for Computational Linguistics

In video data, even if certain frames are lost, viewers can still comprehend the overall content through redundant information (Danier et al., 2024). However, the narrative in yonkoma relies heavily on the close connection between each panel, and any missing information can affect the completeness of the story.

Existing comic datasets such as Sachdeva and Zisserman (2024); Li et al. (2023b); Lee et al. (2021); Aizawa et al. (2020) often lack a structured format, making them difficult to model effectively. In contrast, our YManga dataset offers unique structural advantages: each sample consists of four equally-sized panels, forming a stable 2x2 grid layout. This structure facilitates researchers in exploring the temporal dependencies in visual storytelling and the synergies of multimodal information, which are difficult to achieve in existing comic or image-text datasets.

Based on these characteristics, we design three tasks on the YManga dataset to evaluate the capability of MLLMs in understanding yonkoma data: (1) Panel Sequence Detection (PSD), which tests whether the model can correctly identify the order of yonkoma panels; (2) Intent Generation (IG), which assesses whether the model can generate the author's intended message behind the yonkoma; and (3) Description Generation (DG), which evaluates the model's ability to infer the content of a missing panel when it is masked. These three tasks are illustrated in Figure 1.

To accomplish these tasks, the model must possess three core abilities: accurately recognizing character dialogues, identifying characters across panels, and capturing critical narrative turning points. Through a progressive task design, where we provide yonkoma images, title text, and manual annotations, we observe a considerable gap between the performance of current multimodal models and human understanding. Experiments show that the best model achieves only 67% accuracy on Panel Sequence Detection(PSD), significantly lower than the 90% accuracy achieved by humans. For Intent Generation (IG) and Description Generation (DG), the model's performance heavily depends on the availability of panel descriptions. Even advanced models, such as GPT-4o, struggle to generate accurate intents or descriptions in certain cases, despite being provided with panel descriptions. Our contributions are summarized as follows:

- The YManga dataset we propose is the first dataset specifically curated for the collection of yonkoma-type comics. We collect and filter 1,015 yonkoma strips, manually annotating each with 10 labels, resulting in a total of 10,150 annotations.
- We design three tasks on the YManga dataset and conduct rigorous baseline experiments.
- Through the formulation of five research questions, we perform a comprehensive analysis of the YManga dataset and summarize three main limitations of existing MLLMs.

2 Dataset and Task Setups

We collect yonkoma strips from several influential comic websites, such as Pinterest, GoComics, and from the personal pages of various comic artists. Through a rigorous process of both machine and manual filtering, we retain 1,015 high-quality yonkoma strips. All of these have been authorized by their respective creators. For detailed information on our criteria for filtering data, please refer to Appendix A.

2.1 Task Setups

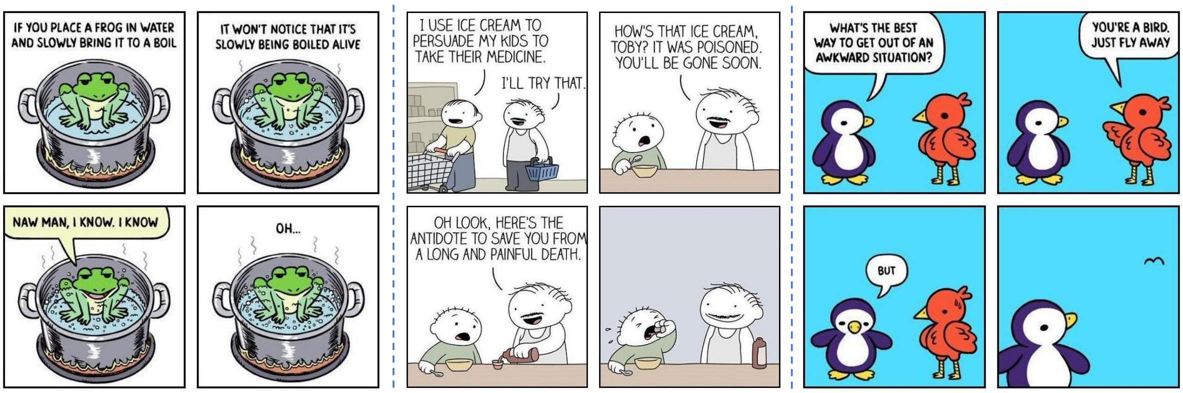
We design three novel tasks on YManga: (1) Panel Sequence Detection(PSD); (2) Intention Generation(IG); (3) Description Generation(DG). The overall statistics of these three tasks are shown in Table 1.

Task 1: Panel Sequence Detection(PSD)

> *Can the models identify whether the sequence of the four panels in the yonkoma is correct?*

Inspired by the next sentence prediction pretraining task proposed by Devlin et al. (2019) in the BERT model, we design this task. We use two methods to swap panels. The first method involves swapping two adjacent panels[3]. The second method involves randomly shuffling all four panels. The sequence of panels in yonkoma is crucial for conveying the coherence and logic of the story. The models need to recognize subtle visual clues within the panels and the semantic relationships between them to determine if the sequence is correct. This requires the models to have not only excellent image processing capabilities but also an understanding of the temporal and causal relationships in the yonkoma's panels.

[3] This method includes three cases: swapping panel 1 and panel 2, panel 2 and panel 3, and panel 3 and panel 4.


IF YOU PLACE A FROG IN WATER AND SLOWLY BRING IT TO A BOIL
IT WON'T NOTICE THAT IT'S SLOWLY BEING BOILED ALIVE
NAW MAN, I KNOW. I KNOW
OH...
I USE ICE CREAM TO PERSUADE MY KIDS TO TAKE THEIR MEDICINE.
I'LL TRY THAT.
HOW'S THAT ICE CREAM, TOBY? IT WAS POISONED. YOU'LL BE GONE SOON.
OH LOOK, HERE'S THE ANTIDOTE TO SAVE YOU FROM A LONG AND PAINFUL DEATH.
WHAT'S THE BEST WAY TO GET OUT OF AN AWKWARD SITUATION?
YOU'RE A BIRD. JUST FLY AWAY
BUT

Title: Tortured Metaphor
Description:
Panel 1: A frog sits in a pot, and a fire is lit underneath. The narrator writes "If you place a frog in water and slowly bring it to a boil."
Panel 2: The frog is still in the pot, and the water begins to bubble slightly. The narrator writes "It won't notice that it's slowly being boiled alive."
Panel 3: The frog says: "Naw man, I know. I know.". The water has begun to boil.
Panel 4: The frog is still in the pot, and the water is boiling. The narrator "Oh..."
Author's Intention: The author describes the "boiling frog in warm water" to satirize those who, even though they are aware of the harm of their environment, still do not take action, causing things to get worse and worse.

Title: Take Medicine
Description:
Panel 1: In the supermarket, character A tells character B that "I use ice cream to persuade my kids to take their medicine.", and character B says "I'll try that."
Panel 2: After returning home, character B tells his son that "How's that ice cream, toby? It was poisoned. You'll be gone soon."
Panel 3: As character B poured the medicine, he continued to say to his son that "Oh look, here's the antidote to save you from a long and painful death."
Panel 4: Character B's son is taking the medicine.
Author's Intention: The author humorously describes the extreme and exaggerated methods parents may use to ensure their children take medication on time, emphasizing the love parents have for their children.

Title: Wing Tip
Description:
Panel 1: A penguin asks a red bird, "What's the best way to get out of an awkward situation?"
Panel 2: The red bird responds, "You're a bird. Just fly away."
Panel 3: The penguin, looking confused, starts to say "BUT ", while the red bird looks at it.
Panel 4: The red bird is gone, and there's a small black line in the sky indicating it has flown away, leaving the penguin standing alone.
Author's Intention: The author humorously illustrates the point that everyone has different advantages and strengths, and when dealing with problems one should not copy other people's solutions, but instead find a method that suits one's own.

Figure 2: For each of the 1,015 yonkoma strips, we annotate all four of their panels. We focus on the characters' dialogue, behaviors, and essential scenes and details. We also distinguish characters that appeared across panels. Additionally, we discuss the authors' intentions in detail and provide accurate manual annotations.

Task 2: Intention Generation(IG)

> *Can the model accurately generate the author's intention in creating the yonkoma?*

We design this task following the general guidelines proposed by Hessel et al. (2023). This task is also known as, can the model understand the sentiment that the yonkoma author want to express when creating the yonkoma? The key to correctly identify the author's intention lies in accurately detecting the turning point of the yonkoma, which typically occurs in the third or fourth panel. To identify this turning point correctly, the models must understand the previous panels and grasp the overall context of the yonkoma. Additionally, it needs to compare the previous panel with the subsequent panel to identify differences, which places extremely high demands on the models' reasoning ability.

Task 3: Description Generation(DG)

> *When a panel is masked, can the models make full use of the existing information to infer what the masked panel should describe?*

Based on the structure of the yonkoma data itself and inspired by the Masked LM pre-training task proposed by Devlin et al. (2019) in the BERT model, we design this task. Since the key elements of a yonkoma often lie in the third and fourth panels, we focus on these two panels. We mask each of these panels separately and provide the author's intention along with the masked yonkoma to see if the models can accurately generate the correct description of the masked panel. We mask the panel using the average color of the masked panel.

For the latter two tasks, we categorize them based on the presence or absence of human-annotated panel descriptions. Specifically, for task IG, the input to the models consists of a yonkoma-title pair, accompanied by the descriptions of four panels when available. For task DG, the input includes a yonkoma with one masked panel, the corresponding title, the author's creative intention, and descriptions of three unmasked panels if they are available. Intuitively, when the description information of the panels are available, the performance of the models will undoubtedly be greatly enhanced. This is indeed the case. When there is panel description information, the output of each model has been greatly improved, which in turn shows that the current multimodal LLMs is facing great challenges in understanding yonkoma strips data.

2.2 Evaluation Metrics

For task PSD, a binary classification task, we use accuracy and F1 score to automatically evaluate the models' predictions. For IG, a generative tasks, we employ automatic evaluation metrics such as Post (2018) + ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020) (using XLNet/xlnet-large-cased), but our primary focus is on human evaluation. We recruit four undergraduate students[4] with no prior experience analyzing our data to assist with the human evaluation, under the guidance of one of the authors of this paper. Specifically, for task IG, we divide the four evaluators into two groups: one group evaluate the predictions with panel descriptions, and the other evaluate the predictions without panel descriptions. The purpose of this division is to ensure that each annotator can only see the correct label once. For each prediction generated by the models, we package it into a prediction-label pair and ask two evaluators to choose which data in the pair is better to determine the generation effect of the models. If both annotator choose the human-annotated text, it indicates that the quality of the text generated by the model is not good enough. Otherwise, it suggests that the model performs well, with no significant difference from human annotation. For task DG, we mainly focus on the following two aspects to observe whether the models' prediction is correct: (1)Consistency: Is the model's prediction consistent with the real panel description? (2)Articulate: Can the model's predictions effectively highlight the author's creative intentions? When all four evaluators agree on these two points, we consider the quality of the prediction to be better. If any of the evaluators disagree with either of these two points, we consider the quality of the prediction to be lower.

2.3 Annotation and Analysis

We recruit three undergraduate students[5] from the School of Languages and Literature to collaborate with one of the authors of this article on manual annotation. See the Appendix B for detailed data annotation guidelines. A random sample of annotations is shown in Figure 2. See Appendix D for more data samples.

	Train	Val	Test
Swap Panel 1,2	1218	406	406
Swap Panel 2,3	1218	406	406
Swap Panel 3,4	1218	406	406
Randomly Shuffle	1218	406	406
Intention Generation	609	203	203
Mask Panel 3	609	203	203
Mask Panel 4	609	203	203

Table 1: Basic statistics of YManga. For PSD, we divide it into four binary classification tasks, each with an equal number of positive and negative examples. For task DG, we divide it into two subtasks: masking the third panel and masking the fourth panel. For these three tasks, we divide the training set, validation set, and test set in a ratio of 3:1:1.

For the creative intention of the yonkoma author, our general annotation standard is to use one sentence to summarize the author's core intention. For each yonkoma, the three annotators engage in heated discussions, reached a consensus, and then perform the manual annotation. The annotation results are reviewed by one of the authors of this paper to form the final annotation. Finally, we create a corpus of 1015 human-annotated author intention explanations with relatively uniform length (mean/median 25.1/25 words, 25.5K total).

For the description of each yonkoma panel, our annotation standards are: (1) For each panel, clearly describe the characters that appear, including their words, expressions, actions, and the background (if important); (2) Accurately distinguish between characters that cross panels and newly appearing characters. The annotators' annotations are also reviewed by the author to form the final annotations. In the end, we complete the annotation of all yonkomas, resulting in a corpus containing 1015 human-annotated yonkoma panel descriptions (mean/median 58.8/58 words, 59.6K total).

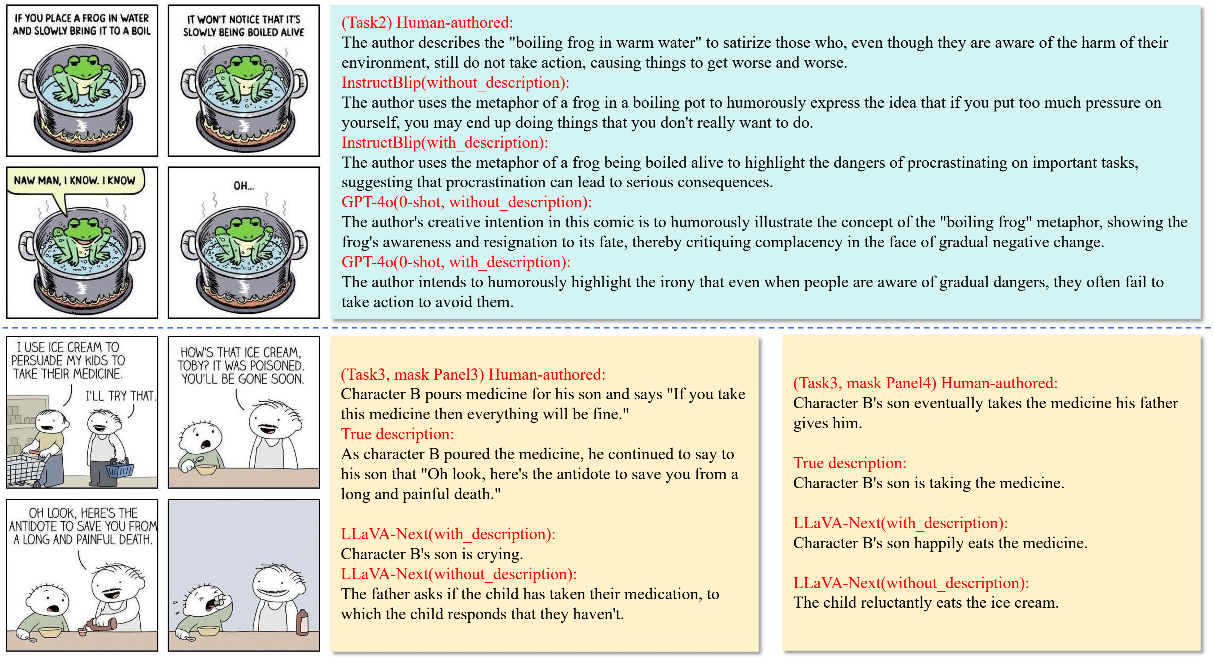
3 Experiments

3.1 Baseline Models

CLIP-ViT. CLIP consists of a text Transformer (Vaswani et al., 2017) and a vision Transformer (Dosovitskiy et al., 2021). We use the vision Transformer of CLIP `ViT-L/14@336px`, which is pre-

[4] We pay each evaluator \$13/hr.

[5] We pay each annotator \$13/hr too.


IF YOU PLACE A FROG IN WATER AND SLOWLY BRING IT TO A BOIL
IT WON'T NOTICE THAT IT'S SLOWLY BEING BOILED ALIVE
NAW MAN, I KNOW. I KNOW
OH...
I USE ICE CREAM TO PERSUADE MY KIDS TO TAKE THEIR MEDICINE.
I'LL TRY THAT.
HOW'S THAT ICE CREAM, TOBY? IT WAS POISONED. YOU'LL BE GONE SOON.
OH LOOK, HERE'S THE ANTIDOTE TO SAVE YOU FROM A LONG AND PAINFUL DEATH.

(Task2) Human-authored:
The author describes the "boiling frog in warm water" to satirize those who, even though they are aware of the harm of their environment, still do not take action, causing things to get worse and worse.
InstructBlip(without_description):
The author uses the metaphor of a frog in a boiling pot to humorously express the idea that if you put too much pressure on yourself, you may end up doing things that you don't really want to do.
InstructBlip(with_description):
The author uses the metaphor of a frog being boiled alive to highlight the dangers of procrastinating on important tasks, suggesting that procrastination can lead to serious consequences.
GPT-4o(0-shot, without_description):
The author's creative intention in this comic is to humorously illustrate the concept of the "boiling frog" metaphor, showing the frog's awareness and resignation to its fate, thereby critiquing complacency in the face of gradual negative change.
GPT-4o(0-shot, with_description):
The author intends to humorously highlight the irony that even when people are aware of gradual dangers, they often fail to take action to avoid them.

(Task3, mask Panel3) Human-authored:
Character B pours medicine for his son and says "If you take this medicine then everything will be fine."
True description:
As character B poured the medicine, he continued to say to his son that "Oh look, here's the antidote to save you from a long and painful death."

LLaVA-Next(with_description):
Character B's son is crying.
LLaVA-Next(without_description):
The father asks if the child has taken their medication, to which the child responds that they haven't.

(Task3, mask Panel4) Human-authored:
Character B's son eventually takes the medicine his father gives him.

True description:
Character B's son is taking the medicine.

LLaVA-Next(with_description):
Character B's son happily eats the medicine.

LLaVA-Next(without_description):
The child reluctantly eats the ice cream.

Figure 3: This is an example of the last two generation tasks. Readers can see more examples in the Appendix D. In this figure, the upper part shows the generation results of Task IG, and the lower part shows the results of the two subtasks of Task DG. It can be clearly seen that (1) the quality of models generation is far from that of human-authored, and (2) the generation result of the models with or without description is also very different.

trained to align images and captions in the WebImageText corpus (400M pairs) (Radford et al., 2021), as a baseline model for PSD, considering only the image information of yonkoma. Specifically, we take the pooling output of CLIP-ViT and subsequently connect an Multilayer Perceptron (Rumelhart et al., 1986) to map it to the binary classification of whether the sequence of yonkoma panels is correct.

InstructBlip. InstructBlip (Dai et al., 2023) consists of a vision Transformer (Dosovitskiy et al., 2021), a Q-Former module, and a large language model. It has undergone vision-language instruction fine-tuning on the pre-trained BLIP-2 (Li et al., 2023a, 2022) model. We choose the `InstructBlip-Flan-T5-XXL`(12B parameters) model which is followed by a `Flan-T5-XXL` (Raffel et al., 2020; Chung et al., 2022). For task PSD, we use its vision Transformer and Q-Former module to encode the image-text data of yonkoma, obtain its pooling output, and then connect an Multilayer Perceptron (Rumelhart et al., 1986) to map it to the binary classification of whether the sequence of the yonkoma panels is correct. For Tasks 2 and 3, we fine-tune `InstructBlip-Flan-T5-XXL` using Low-Rank Adaptation (Hu et al., 2022).

LLaVA-NeXT. LLaVA-NeXT (Liu et al., 2024) is a multimodal large language model improved upon LLaVA (Liu et al., 2023b,a). It has enhanced reasoning abilities, optical character recognition (OCR), and world knowledge, outperforming many multimodal LLMs on various general visual language tasks, including Gemini Pro (Anil et al., 2023). We choose `llava-v1.6-mistral-7b-hf`(7B parameters), an open-source model available on Hugging Face, as the baseline for our last two generation tasks. We also use Low-Rank Adaptation (Hu et al., 2022) to fine-tune it.

GPT-4o. We also test our three tasks on the GPT-4o (OpenAI, 2024a,b). We design detailed prompts for these three tasks. For task PSD, we set the model to output "Yes" or "No" to achieve the effect of binary classification. For the next two generative tasks, we design prompts for the presence or absence of panel descriptions. See Appendix C for more prompt details.

3.2 Fine-tuning Details

All of our locally run models are trained on a single A100 GPU using the PyTorch framework (Paszke et al., 2019). We select the AdamW optimizer (Loshchilov and Hutter, 2019). The specific parameters use for the AdamW optimizer are as follows: the learning rate is set to $1e-5$, with β_1 and β_2 val-

Model	Swap Panel 1, 2		Swap Panel 2, 3		Swap Panel 3, 4		Randomly Shuffle	
	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Random	50	50	50	50	50	50	50	50
CLIP-ViT	58.54	60.65	60.24	60.91	58.29	59.57	57.80	50.14
Q-Former	60.73	63.16	**65.61**	**62.99**	**67.32**	**67.48**	62.20	**65.63**
LLaVA-NeXT	59.35	59.85	59.85	60.58	60.49	58.04	61.42	59.47
GPT-4o(0-shot)	**63.66**	**63.39**	58.05	54.01	57.80	56.64	**64.79**	63.26
Human	92.80	92.80	90.34	91.77	90.24	90.20	95.24	90.91

Table 2: Results of a baseline experiment to detect if the sequence of yonkoma panels is correct. Q-Former is the vision Transformer and Q-Former module of the `instructblip-flan-t5-xxl` mentioned in Section 3.1. We take the average of 5 cross-validation splits. During training and testing, the CLIP-ViT model can only access the image information of the yonkoma, while other models and human evaluation can access the image and title text information of the yonkoma. We make four settings for the sequence of panels, namely swapping panels 1, 2; swapping panels 2, 3; swapping panels 3, 4; and randomly shuffling the panels.

ues of 0.9 and 0.999, respectively, and a weight decay rate of 0.01. For each individual task, we train the models until they reach convergence, ensuring optimal performance before proceeding to the testing phase. To achieve reliable and unbiased results, we employ a 5-fold cross-validation approach. This method allowed us to utilize every sample in the dataset for both training and testing, thereby enhancing the robustness and generalizability of our models. Specifically, for the task of PSD, we conduct comprehensive parameter fine-tuning on the classification models. This involve adjusting all model parameters to achieve the best performance. For task IG and DG, we limit the fine-tuning process to only the linear layers of the models. This selective fine-tuning strategy help in maintaining computational efficiency while still achieving satisfactory performance levels. We divide YManga into 5 cross-validation groups, and take the average of these five cross-validation groups as our experimental results.

3.3 Panel Sequence Detection

Table 2 shows the experimental results of PSD. As can be seen, our fine-tuned Q-Former module achieved slightly better results than the 0-shot GPT-4o model, but all our models have a large gap with human evaluation. In order to conduct in-depth analysis and exploration of the experimental results of PSD, we design two research questions(RQs).

RQ1: What does the results of Q-Former and GPT-4o show? The main difference between the fine-tuned Q-Former and the zero-shot GPT-4o is their recognition of the exchange between panel 2 and 3, and between panel 3 and 4. It can be seen that for these two subtasks, the zero-shot GPT-4o performs worse than even the CLIP-ViT model, which lacks title information. We believe this is due to the features of yonkoma data. Panel 1 must introduce the background information of the story, with all the information appearing for the first time. The subsequent panels build on this basic information. Therefore, even with zero-shot GPT-4o, it performs better than the trained model in recognizing the sequence of the first two panels or randomly shuffling the sequence. However, for the sequence of the last three panels, although the visual reasoning ability of GPT-4o is unquestionable, it still does not grasp the details of the yonkoma images as well as the smaller model fine-tuned on YManga.

RQ2: What features of yonkoma data does human evaluation reveal? How does it differ from model testing? As for the results of human evaluation, we can see that the trend is very similar to that of GPT-4o. The performance on swapping Panel 1 and Panel 2, and on randomly shuffling, is slightly better than the other two subtasks. This undoubtedly reveals a feature of yonkoma data. To some extent, the sequence of the last three panels of some yonkoma may have an implicit relationship that is not very obvious. For example, in the third example of Figure 2, even if we swap Panel 2 and Panel 3, it does not significantly impact the development and expression of the whole story. However, this type of yonkoma data only accounts for a small portion. In most cases, humans can easily determine whether the sequence of panels is correct. We did not filter out this type of data. On the contrary, we believe that the presence of this type of data makes our model more robust, less prone to overfitting, and beneficial for our subsequent research.

3.4 Intention Generation

Table 3 presents the automatic evaluation and the human evaluation results of the model generation of task IG. We show some examples generated

Model	Task 2			Task 3 Mask Panel 3	Task 3 Mask Panel 4
w/o desc	ROUGE	Bert-Score	Human	Human	
InstructBlip	26.15	52.40	37%	44%	41%
LLaVA-NeXT(0-shot)	26.42	52.23	35%	41%	36%
LLaVA-NeXT(finetuned)	26.55	51.87	33%	43%	33%
GPT-4o(0-shot)	**28.73**	54.51	53%	54%	49%
GPT-4o(5-shot)	27.65	**54.69**	**69%**	**68%**	**54%**
w/ desc					
InstructBlip	**28.67**	**55.47**	48%	63%	48%
LLaVA-NeXT(0-shot)	27.59	53.48	44%	55%	42%
LLaVA-NeXT(finetuned)	27.98	53.76	45%	57%	42%
GPT-4o(0-shot)	28.51	54.82	60%	60%	52%
GPT-4o(5-shot)	28.39	55.12	**74%**	**73%**	**62%**

Table 3: Here are the results of several baseline models on the IG and DG tasks. For the IG task, we primarily used a combination of automatic evaluation metrics and human evaluation. For the DG task, since it involves open-domain text generation, we mainly relied on human evaluation. The percentage in the human evaluation results refers to the proportion of high-quality outputs generated by the model.

by the model in Figure 3. For specific evaluation methods, see Section 2.2, Evaluation Metrics. The prompt of GPT-4o is shown in the Appendix C. We also prepare two Research Questions to analyze and explore this task.

RQ3: What does it reveal that the generated results with description information are better than those without description information? We compare the impact of having description information on the generation results of Task 2 and Task 3. The MLLMs performed better when provided with description information, which highlights their limited ability to understand multimodal information. When the models have access to panel descriptions, their performance improve significantly, consistent with the role of dense information as proposed by Fan et al. (2024). However, for more advanced models, such as GPT-4o, the improvement was relatively minor. This clearly indicates that current models still face substantial challenges in accurately understanding yonkoma data. It also demonstrates that Task 1 serves as a foundation for Task 2 and Task 3.

RQ4: For GPT-4o, what does it mean that the 5-shot result is better than the 0-shot result? Brown et al. (2020) first introduce in-context learning as a special form of prompting. We believe that for generative tasks, this approach is somewhat analogous to human alignment (Ouyang et al., 2022; Ziegler et al., 2019), where large models learn human language habits and then imitate the tone to express answers to questions. As shown by the experimental results in Table 3, while this imitation improves the model's expressive abilities, we argue that MLLMs are primarily imitating human output patterns and do not possess a strong understanding of "emotion".

3.5 Description Generation

Table 3 shows the experimental results of task DG. We only perform manual evaluations on this open-domain task. Table 3 presents the manual evaluation results generated by the model. These percentages represent the proportion of high-quality descriptions generated out of the total number of generated descriptions.For task DG, we raise one research question to analyze and explore this task.

RQ5: Why is the model better at generating descriptions for masked panel 3 than for masked panel 4? The DG task tests the reasoning ability of MLLMs in open domains. After discussion, we identify two reasons for this phenomenon. The first reason is that when generating the description for the third panel, we can refer not only to the author's intent and the previous panels but also to the information from the fourth panel. This approach helps anchor the generation of the third panel's description to some extent. However, when generating the description for the fourth panel, the model lacks this advantage and is more prone to hallucination, as noted by Dhuliawala et al. (2024); Qu et al. (2024). The second reason relates to the characteristics of yonkoma data itself. Typically, a turning point appears in the third panel, which is closely tied to the emotions the author aims to convey, making it easier for the model to learn. This also highlights the weaker capability of MLLMs in generating text in more open-ended environments.

We formulate five independent research questions to thoroughly and comprehensively explore the limitations of MLLMs. First, MLLMs are less effective in capturing image details compared to smaller models that perform classification based on feature extraction in task 1. Second, when dealing with complex image structures, such as yonkoma, MLLMs are more prone to hallucinations in cross-modal tasks. Third, MLLMs also show limitations in their generative capabilities in open-domain. These shortcomings hinder the further development of MLLMs.

4 Related Work

Research on comics analysis. Comics, as a highly comprehensive form of multimodal data, have attracted increasing attention from researchers in recent years. However, existing research on comics mainly focuses on specific aspects. For instance, Martínek et al. (2024) explore dialogue recognition in images through comics and introduced a dataset containing 1,438 annotated panels. Kovanen and Aizawa (2015) propose a hierarchical method to study the reading order of comic text bubbles. Hinami et al. (2021) incorporate contextual information extracted from comics into a translation system, improving translation accuracy. Agrawal et al. (2023) focus on comic character dialogue generation and contribute a new dataset called COMSET. Kim et al. (2024) leverage the sequential features of comic data to solve the description generation problem for related data. He et al. (2018) propose a task to extract irregular comic panels, while Xu et al. (2023) segment comic panels and propose a Panel-Page-Aware comic type classification model. He et al. (2017) introduce the SReN model for detecting panels in comics. Iyyer et al. (2017) presented a comic dataset called COMICS and introduced three cloze-style tasks, requiring the model to predict the narrative of a panel given the context of the previous n panels. Baek et al. (2022) construct a new Japanese comic onomatopoeia dataset called COO, which challenges the recognition of irregular text. Although there is a variety of work on comic data, these studies focus on specific features of comics and fail to fully leverage the comprehensive nature of comic data. Moreover, these works are not suitable for testing MLLMs, which possess strong multimodal capabilities.

The development of multimodal LLMs. The successful application of large language models has promoted the development of research in the multimodal field and paved the way for the construction of multimodal large language models. There are three main methods for building multimodal large-scale language-based models, each aiming to achieve strong zero-shot generalization capabilities in the field of visual language. For example, Flamingo (Alayrac et al., 2022) is a pioneer in this field, using frozen visual encoders and large language models equipped with gated cross-attention for cross-modal alignment. This method has been widely adopted by models such as LLaVA (Liu et al., 2023b) and Shikra (Chen et al., 2023). However, a significant limitation of this approach is the creation of lengthy visual sequences. To address this issue, BLIP-2 (Li et al., 2023a) drew inspiration from DETR (Carion et al., 2020) and developed a Q-former to effectively reduce the sequence length of visual features. Kosmos-1 (Huang et al., 2023), mPLUG-Owl2 (Ye et al., 2023), and MiniGPT-4 (Zhu et al., 2023) all adopt this design to reduce the visual sequences. Despite the rapid development of MLLMs, our experiments show that these models have limitations in handling complex image-text modal reasoning problems.

5 Conclusion and Future Work

To explore the limitations of MLLMs, we introduce the YManga dataset, which consists of 1,015 high-quality Yonkoma Manga samples with corresponding human annotations. We designed three tasks based on YManga: Panel Sequence Detection (PSD), Intent Generation (IG), and Description Generation for Masked Panels (DG). Through rigorous baseline experiments and in-depth analysis of the results, we identified three main limitations of MLLMs: insufficient ability to capture fine-grained image details, a tendency to hallucinate when handling complex image data, and inadequate generative capabilities in open-domain tasks.

Our future work will build on the YManga dataset to further explore ways to mitigate the three limitations of MLLMs. This will primarily focus on two areas: first, designing more efficient methods for modal alignment to capture more image details, and second, improving the generative quality of the models by incorporating external knowledge bases or leveraging techniques such as reinforcement learning.

6 Limitations

In this work, we collect and annotate Yonkoma yonkoma data, organize it into a dataset, and conduct rigorous experiments on this dataset. However, several limitations require further exploration. First, the purpose of proposing YManga is to test the model's ability to understand multimodal data, so we do not specifically design a model to explore the upper limit of performance on these tasks. Nonetheless, this constitutes a limitation of this work. Secondly, although the task of understanding yonkoma data is very comprehensive, a model's weak performance in this area does not necessarily indicate its overall incompetence. We design this task to explore one aspect of the model's capabilities, but it should not be considered the most important criterion for evaluating the model.

7 Acknowledgements

We thank reviewers for their comments, which provided some insights on this research that will further influence our future work. This work is partially supported by grants from the Key R&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Natural Science Foundation of China award numbers (62376051, 62366040, 62076046, 62066044, 61976036, 61702080) and the Fundamental Research Funds for the Central Universities award number (DUT24LAB123).

References

Harsh Agrawal, Aditya Mishra, Manish Gupta, and Mausam. 2023. Multimodal persona based generation of comic dialogs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14150–14164. Association for Computational Linguistics.

Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. 2020. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE Multim.*, 27(2):8–18.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. 2022. COO: comic onomatopoeia dataset for recognizing arbitrary or truncated texts. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 267–283. Springer.

Gary Bradski. 2000. The opencv library. In *Dr. Dobb's Journal of Software Tools*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, and et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with 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*.

Duolikun Danier, Fan Zhang, and David R. Bull. 2024. LDMVFI: video frame interpolation with latent diffusion models. 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 1472–1480. AAAI Press.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhiyuan Fan, Zhihong Chen, and Benyou Wang. 2024. Exploring the potential of dense information in multimodal alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13440–13451, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Rafael C. Gonzalez and Richard E. Woods. 2018. *Digital Image Processing*, 4th edition. Pearson.

Zheqi He, Yafeng Zhou, Yongtao Wang, and Zhi Tang. 2017. Sren: Shape regression network for comic storyboard extraction. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4937–4938. AAAI Press.

Zheqi He, Yafeng Zhou, Yongtao Wang, Siwei Wang, Xiaoqing Lu, Zhi Tang, and Ling Cai. 2018. An end-to-end quadrilateral regression network for comic panel extraction. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 887–895. ACM.

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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.

Ryota Hinami, Shonosuke Ishiwatari, Kazuhiko Yasuda, and Yusuke Matsui. 2021. Towards fully automated manga translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12998–13008. AAAI Press.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L. Boyd-Graber, Hal Daumé III, and Larry S. Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6478–6487. IEEE Computer Society.

Suhyun Kim, Semin Lee, Kyungok Kim, and Uran Oh. 2024. Utilizing a dense video captioning technique for generating image descriptions of comics for people with visual impairments. In *Proceedings of the 29th International Conference on Intelligent User Interfaces, IUI 2024, Greenville, SC, USA, March 18-21, 2024*, pages 750–760. ACM.

Samu Kovanen and Kiyoharu Aizawa. 2015. A layered method for determining manga text bubble reading order. In *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, pages 4283–4287. IEEE.

Yunjung Lee, Hwayeon Joh, Suhyeon Yoo, and Uran Oh. 2021. Accesscomics: an accessible digital comic book reader for people with visual impairments. In *W4A '21: 18th Web for All Conference, Virtual Event / Ljubljana, Slovenia, April 19-20, 2021*, pages 2:1–2:11. ACM.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yingxuan Li, Kiyoharu Aizawa, and Yusuke Matsui. 2023b. Manga109dialog A large-scale dialogue dataset for comics speaker detection. *CoRR*, abs/2306.17469.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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

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

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Jirí Martínek, Pavel Král, Ladislav Lenc, and Josef Baloun. 2024. COMICORDA: dialogue act recognition in comic books. 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 3566–3578. ELRA and ICCL.

OpenAI. 2024a. Ai and covert influence operations: Latest trends. Technical report, OpenAI.

OpenAI. 2024b. Hello gpt-4o.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Ge Qu, Jinyang Li, Bowen Li, Bowen Qin, Nan Huo, Chenhao Ma, and Reynold Cheng. 2024. Before generation, align it! a novel and effective strategy for mitigating hallucinations in text-to-SQL generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5456–5471, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Ragav Sachdeva and Andrew Zisserman. 2024. The manga whisperer: Automatically generating transcriptions for comics. *CoRR*, abs/2401.10224.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Chenshu Xu, Xuemiao Xu, Nanxuan Zhao, Weiwei Cai, Huaidong Zhang, Chengze Li, and Xueting Liu. 2023. Panel-page-aware comic genre understanding. *IEEE Trans. Image Process.*, 32:2636–2648.

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. *CoRR*, abs/2311.04257.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *CoRR*, abs/1909.08593.

A Data Collection Standards

We follow general data collection standards (Hessel et al., 2023) while also considering the unique characteristics of the Yonkoma Manga dataset. Our data collection process involve a two-step review and filtering procedure, consisting of machine-based filtering followed by manual filtering.

A.1 Machine-based Collection Criteria

Filtering out low-resolution Yonkoma Manga:

- We use Python's OpenCV (Bradski, 2000) library to read images, extracting their DPI (dots per inch) and filtering out Yonkoma Manga with a DPI lower than 96. This ensures that key elements such as characters' facial expressions, actions, dialogue text, and narration are clearly visible.
- Bilateral and median filters are applied to the remaining images to remove Gaussian noise and salt-and-pepper noise (Gonzalez and Woods, 2018).
- Canny edge detection (Canny, 1986) is employed to compute the proportion of edge pixels. Yonkoma Manga with less than 10% edge pixels are filtered out.

Filtering out Yonkoma Manga with inconsistent panel sizes:

- The `findContours` function in OpenCV (Bradski, 2000) is used to extract the contours of the four panels.
- The aspect ratio and area of each panel are calculated.
- Yonkoma Manga with more than a 5% difference in aspect ratio or area between any two panels are filtered out.

Filtering out duplicate Yonkoma Manga:

- The SIFT (Scale-Invariant Feature Transform) (Lowe, 2004) algorithm is used to compare image fingerprints.
- A fingerprint database is established to filter out Yonkoma Manga with a similarity score greater than 0.75.
- Yonkoma Manga with a similarity score between 0.5 and 0.75 are subjected to manual review for further filtering.

A.2 Manual Collection Criteria

Filtering out Yonkoma Manga with ambiguous emotional expressions:

- Yonkoma Manga containing ambiguous language or behavior are filtered out.
- Yonkoma Manga with self-contradictory content or inconsistent semantics are filtered out.
- Yonkoma Manga with multiple possible interpretations, metaphors, or implicit content are filtered out.
- Yonkoma Manga containing elements that may result in cultural misunderstanding or cognitive bias are filtered out.

Filtering out Yonkoma Manga that conflict with mainstream international values:

- Yonkoma Manga containing offensive depictions of violence, pornography, gore, or illegal activity are filtered out.
- Yonkoma Manga promoting racism, colorism, gender discrimination, or ableist biases are filtered out.
- Yonkoma Manga potentially violating intellectual property, infringing on privacy, or posing security risks are filtered out.
- Yonkoma Manga containing religious or politically sensitive topics are filtered out.
- Yonkoma Manga that violate fundamental human rights and dignity are filtered out.
- Yonkoma Manga promoting pseudoscience, anti-intellectualism, or fringe theories are filtered out.
- Yonkoma Manga containing dangerous rhetoric or misinformation are filtered out.

B Data Annotation Standards

We follow general data annotation standards (Hessel et al., 2023) while also taking into account the unique characteristics of the Yonkoma Manga dataset. The annotation process is divided into two categories: Yonkoma Manga description annotation and emotional expression annotation.

B.1 Annotation Method

Below are the specific guidelines for Yonkoma Manga description annotation:

Character Identification

- Assign a unique identifier to each character appearing (e.g., A, B, C, etc.).
- Describe each character's physical attributes (e.g., gender, age, clothing, etc.).
- Track the continuity of the characters across different panels.

Dialogue Annotation

- Accurately record the content of each character's dialogue.
- Identify the speaker for each dialogue segment.
- Annotate the tone or emotion of the dialogue (e.g., anger, surprise, happiness, etc.).

Scene Description

- Describe the background environment (e.g., indoor/outdoor, specific locations, etc.).
- Record important objects or elements in the scene.
- Note any changes in the scene over time.

Action Description

- Provide detailed descriptions of character actions and facial expressions.
- Pay attention to the sequence and causality of actions.

Visual Effects

- Record special visual effects (e.g., close-up shots, motion lines, etc.).
- Pay attention to the use of color and compositional features.

Below are the guidelines for annotating the emotional expression of the author:

- Summarize the core creative intent of the author in one sentence.
- Consider the overall theme and emotional tone of the Yonkoma Manga.
- Analyze the turning point (usually found in the third or fourth panel).
- Consider the humorous elements and any satirical undertones in the Yonkoma Manga.

B.2 Quality Review

For the quality standards of manual annotations, we use a combination of machine and human evaluation. To ensure high-quality annotations, we adapt the BERT-score metric to evaluate the agreement rate among different annotators.

For the panel description annotations, our agreement rate requirements are as follows:

- Character Identification: ≥ 0.9
- Dialogue Annotation: ≥ 0.95 (due to the importance of dialogue)
- Scene Description: ≥ 0.85
- Action Description: ≥ 0.85
- Visual Effects: ≥ 0.8

For the emotional expression annotations, our agreement rate requirements are:

- Overall Agreement Rate: ≥ 0.8
- Core Intent Agreement Rate: ≥ 0.9 (even if expressed differently, the core meaning should remain consistent)

If the agreement rate for a certain type of annotation falls below the required threshold, we will arrange additional training for the annotators. For samples with particularly low agreement rates, a third-party annotator will be brought in to arbitrate.

C Prompts of GPT-4o

In this section we show the prompts used by GPT-4o for all tasks.

- This is the prompt of PSD.

<image>You are an assistant who is good at reading four-panel comics.
The title of this four-panel comic is <Title>.
Please tell me whether the sequence of the four panels of this comic is correct.
You only need to answer "Yes." or "No."

Figure 4: This is the prompt for PSD. We only used 0-shot for PSD.

- This is the 0-shot prompt of IG.

<image> You are an assistant who is good at reading four-panel comics and can accurately grasp the author's emotions.
The title of this four-panel comic is <Title>.
<Description of panels if with description.>
Please tell me in one sentence what the author of this comic created to express.

- This is the 5-shot prompt of IG.

You are an assistant who is good at reading four-panel comics and can accurately grasp the author's emotions.
<Sample 1>,<Sample 2>,<Sample 3>,<Sample 4>,<Sample 5>
<image> The title of this four-frame comic is <Title>.
<Description of panels if with description.>
Please tell me in one sentence what the author of this comic created to express.

Figure 5: These are the four prompts for IG, with and without descriptions for the 0-shot and 5-shot cases.

- This is the 0-shot prompt of DG.

<image>You are an assistant who is good at reading four-panel comics and can infer the contents of masked panel.
The title of this four-panel comic is <Title>, and now its panel <3/4> is masked.
<Description of panels if with description.>
Please tell me in one sentence what the masked panel should describe.

- This is the 5-shot prompt of DG.

You are an assistant who is good at reading four-panel comics and can infer the contents of masked panel.
<Sample 1>,<Sample 2>,<Sample 3>,<Sample 4>,<Sample 5>
<image> The title of this four-panel comic is Title, and now its panel 3/4 is masked.
<Description of panels if with description.>
The panel 3/4 is masked.
Please tell me in one sentence what the masked panel should describe.

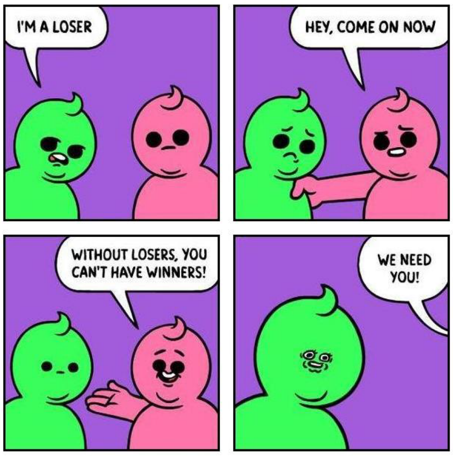
Figure 6: These are the four prompts for DG, with and without descriptions for the 0-shot and 5-shot cases.

Prompt	Accuracy	F1-Socre
Original Prompt	64.79	63.36
"Yes,no" ->"1,0"	62.84	62.35
"Yes,no" ->"True,False"	65.46	64.04
Swap sentence 1 and 2	64.86	64.49
Swap sentence 2 and 3	64.37	65.54

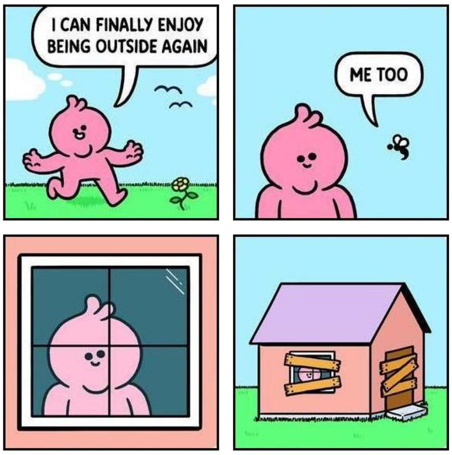
Table 4: Prompt Sensitivity of GPT-4o

We design detailed prompts for all three tasks to ensure precise and effective results. For PSD, we specifically structure the prompt to force the model to output either "Yes." or "No." This approach helps us achieve the effect of binary classification, simplifying the decision-making process. For IG and DG, we take a different approach by designing prompts that focus on the presence or absence of panel descriptions. This method allows us to guide the model's generative capabilities more effectively, ensuring that the output aligns with our specific requirements and expectations.Finally, we test the issue of prompt sensitivity of GPT-4o, as shown in Table 4.

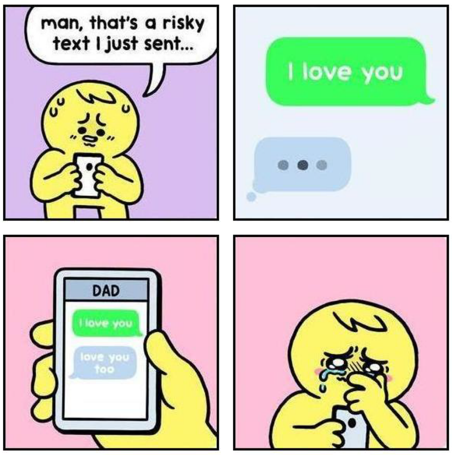
D Data Samples



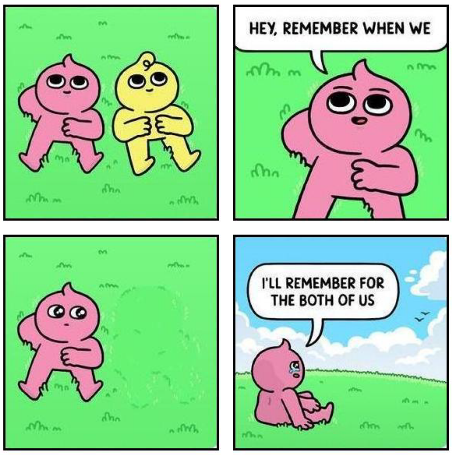
Loser



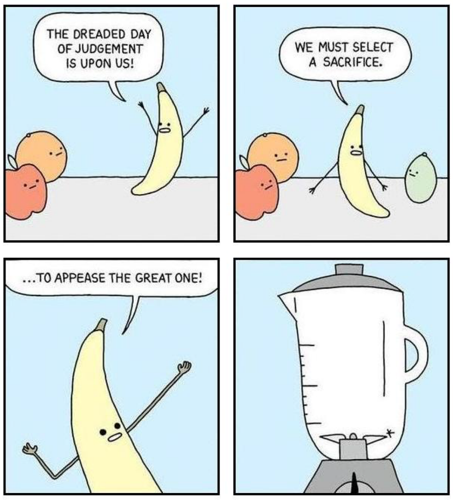
Spring Chicken



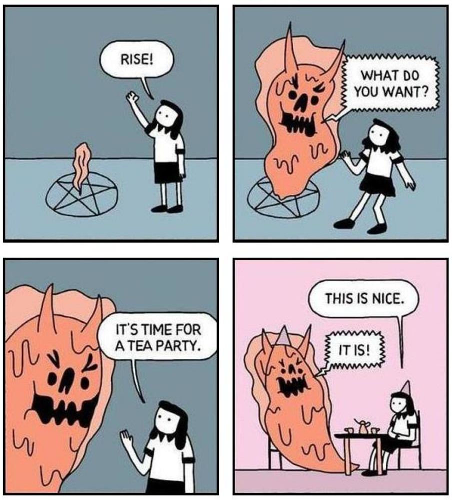
Risky Text



Remember


THE DREADED DAY OF JUDGEMENT IS UPON US!
WE MUST SELECT A SACRIFICE.
...TO APPEASE THE GREAT ONE!

the Great One



Uttuku Risk