

From Panels to Prose: Generating Literary Narratives from Comics

Supplementary Material

A. Distinctive contributions of the Magiv3 model

While the main contribution of this work is the pipeline we propose as a whole, along with the dataset annotations which permit future research and benchmarking, in this section we want to draw attention to the proposed Magiv3 model. Since this model is built on top of Florence-2 [43], we want to be clear and transparent about how our model is distinctive from prior works. Tab. 5 provides an overview.

	Detection	Character Clustering	Text - Character Matching	OCR	Captioning	Character Grounding
DASS [36]	Characters	No	No	No	No	No
Magiv1 [30]	Characters, Texts, Panels	Yes, using a separate character crop encoder	Yes	No, uses a separate model	No	No
Magiv2 [31]	Characters, Texts, Panels, Tails	Yes, using a separate character crop encoder	Yes	No, uses a separate model	No	No
Florence-2 [43]	Zero-shot, performs poorly	No	No	Yes, but line-by-line instead of text-blocks	Yes, low quality	Not really
Magiv3 [Ours]	Characters, Texts, Panels, Tails	Yes	Yes	Yes	No	Yes

Table 5. Comparison of Magiv3 with prior works.

The notable advantage of Magiv3 is that it is a unified architecture for many manga-related tasks with a single set of model weights. Despite being built on top of Florence-2, Magiv3’s abilities are significantly different. For instance, as shown in Fig. 7, Florence-2 cannot meaningfully ground characters in the given captions. It has no notion of context specific phrase grounding and ultimately finds “class labels” from the captions and predicts boxes for them. Florence-2 also cannot perform character clustering or speaker diarisation which Magiv3 can. Conversely, Magiv3 has not been trained to generate captions due to the high training costs required, while Florence-2 can generate [low quality] captions.

Additionally, Magiv3 can do everything Magiv1, Magiv2 etc. can (and more), without needing a separate crop encoder trained contrastively, or a separate OCR model, or post-processing for ordering. In fact, due to the next token prediction paradigm, we do not have to tweak any thresholds to filter out low confidence predictions; unlike prior works which use finicky hyperparameters which are sensitive to near-boundary predictions. As shown in Fig. 8, the optimal threshold for character detection in Magiv2 tends to either miss predictions or add false positives. In contrast, Magiv3 works much better and hence has a much higher precision for character detection (as noted in the main paper).

The model is able to handle non-standard grid layouts. For example, in Fig. 8 there is a “non standard” case with a borderless panel and the layout isn’t a typical grid. In general, the model is extremely robust to various panel arrangements. We include a few more examples in Fig. 9.

Finally, in terms of generalisation, even though we exclusively train Magiv3 on Japanese comics, it works surprisingly well on Western comics as well. We include examples of this in Fig. 10, where we show predictions on Western comics from the Golden Age of Comic Books, which were not part of the training data. Beyond English texts, Magiv3 can still generalize to a degree—for example, it can successfully predict panels, detect and cluster characters for Chinese/Korean comics etc. However, it has poor text detection and no OCR performance on these languages as one might expect.

B. More on character grounding

As mentioned in Sec. 4.2 of the main paper, here we explore an alternative, training-free, and zero-shot method for obtaining pre-grounded captions for manga panels. In fact this method is leveraged to acquire training data for Magiv3, as alluded to in Sec. 7 of the main paper. Finally, we discuss the inherent limitations of this technique.

The core idea behind this approach is simple yet effective. First, we detect all characters in a manga page using an existing model, thereby identifying the bounding boxes of all characters within individual panels. This ensures that the spatial locations of the characters are already well-defined. Subsequently, inspired by the visual prompting methodology in [44], we overlay the detected bounding boxes on the panel image and assign each bounding box a unique identifier. The modified panel, now augmented with visual prompts and unique IDs, is fed into GPT-4o-2024-08-06 alongside a carefully crafted instruction prompt.

In this prompt, we explicitly instruct GPT-4o-2024-08-06 to pre-ground the characters by generating captions that include character-referring phrases in a predefined format that links the overlaid bounding box ID to the character-referring phrase. This ensures the output is structured and ready for automated parsing. An example is illustrated in Fig. 11. Once GPT-4o-

Caption

The image depicts a manga-style illustration with two characters in a school setting. The first character, a girl with short hair, is wearing a school uniform consisting of a skirt and a patterned top. She is pointing towards the background, indicating urgency or surprise. The second character, also a girl in a similar uniform, is standing near a corner and appears to be reacting to the first character. The environment suggests a school hallway or courtyard, with a fence and windows along a building wall. The scene conveys a sense of movement and surprise, with the characters seemingly in motion or reacting to something happening off-panel.

Florence2

Grounding Referring Expressions

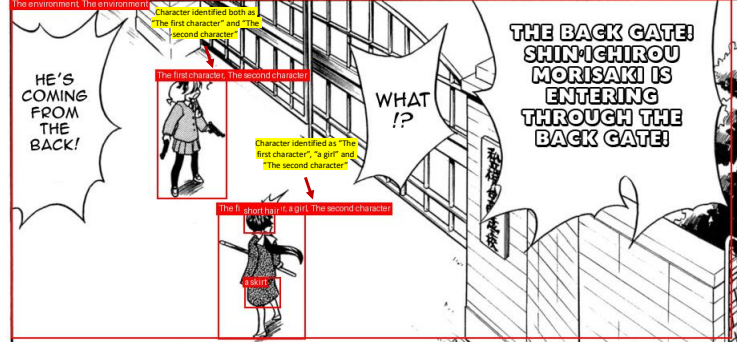


Image: Arisa

Figure 7. Florence2 Character Grounding. Florence2 is trained to perform the *caption to phrase grounding* task. For this task, the input to the model is an image, along with its caption, and the model outputs a set of [phrase, bbox] pairs. We visualise these predicted phrases and the corresponding boxes (right). As noted in the text highlighted in yellow, Florence2 grounds the phrase “The first character” to two different characters, which is clearly a contradiction. This contradiction is also apparent for the phrase “The second character”. This shows that Florence2’s *caption to phrase grounding* ability is only adept at extracting “class labels” from the captions (e.g. character) and predicting boxes for all locations where this class object appears. This behaviour is inadequate for the character grounding task proposed in this work where the model should disambiguate character referring phrases based on the context.

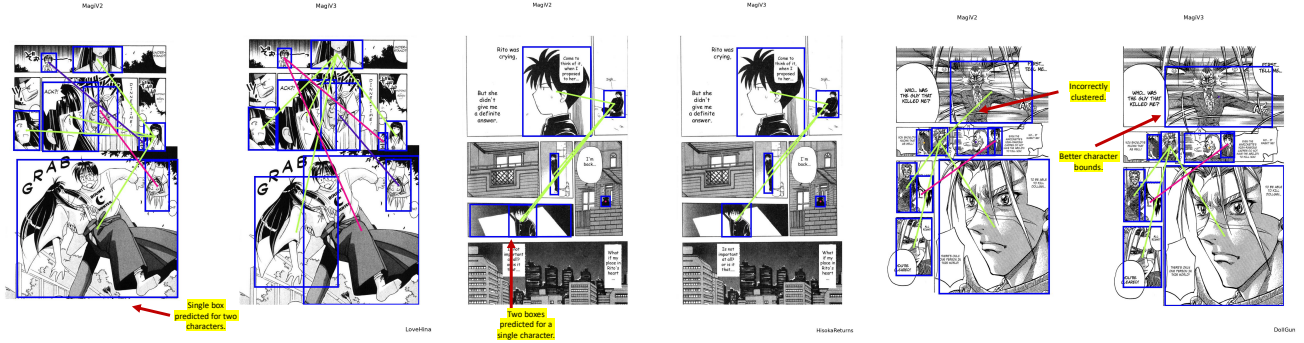


Figure 8. Magiv2 vs Magiv3 character detection and clustering. The predicted character boxes are highlighted in blue. The connected components formed by the colourful lines show the predicted character clusters. As noted in the text in yellow, Magiv2 can often miss predictions (grouping multiple characters in a single box) or over predict (predicting multiple boxes for a single character). This issue is less apparent in Magiv3, while also providing tighter boundaries for characters.

2024-08-06 generates the captions, we parse the output according to the predefined format, extracting the grounded character annotations. This process allows us to effectively mine character grounding information in a zero-shot manner.

To generate training data for Magiv3, we apply this method to panels from the PopManga dev set. The annotations mined through this pipeline are then used to train Magiv3.

Limitations: There are a few limitations of this method, which are thus reflected in Magiv3’s abilities. First, there is a very obvious and significant loss in caption quality, compared to directly asking GPT-4o-2024-08-06 to caption the raw panel image. This is likely because the interleaved brackets are unnatural for typical captions and negatively condition future tokens. Second, the model struggles in crowded scenes where multiple characters are present in close proximity or overlapping. Overlaying the bounding boxes on such panels biases GPT-4o-2024-08-06’s interpretation by making all characters appear equally important, even if many are irrelevant. This often leads the model to describe every character in detail, resulting in verbose and less informative captions. Without the bounding boxes, the model is typically more effective at identifying the central focus of the panel and ignoring minor or background characters. However, the bounding boxes disrupt this natural prioritisation, deteriorating the quality of the captions and their utility for training data generation. Additionally, in densely packed panels with many characters, GPT-4o-2024-08-06 struggles to resolve ambiguities, leading to incorrect IDs matched to referring phrases. Finally, GPT-4o-2024-08-06 occasionally fails to adhere to the predefined output format, resulting in parsing errors and necessitating manual intervention or re-computation. These issues collectively make this approach both noisy and resource-intensive, particularly when scaling to larger datasets.

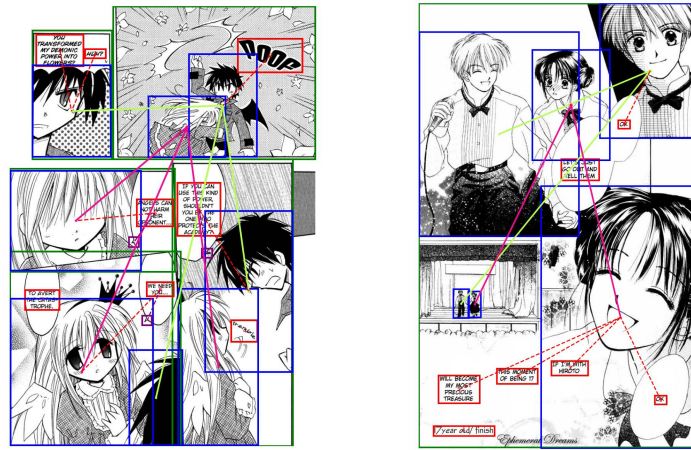


Figure 9. Non Standard Panel Arrangements.

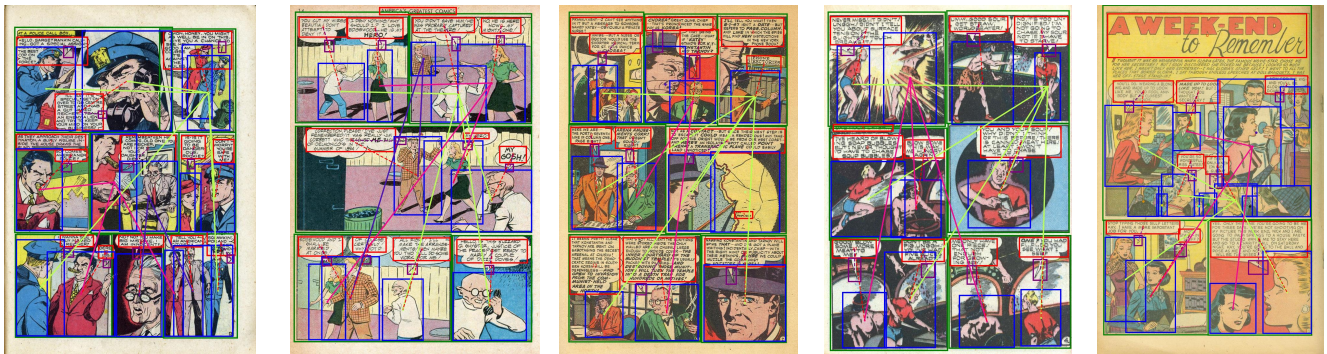
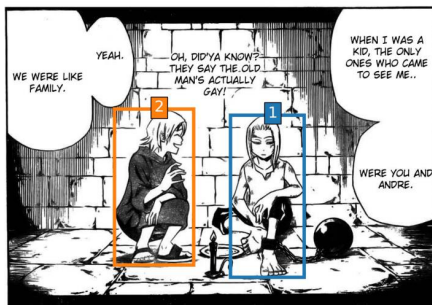


Figure 10. **Magiv3's predictions on Western Comics.** The predicted panels bboxes are in green, characters in blue, text boxes in red and speech bubble tails in purple. The character clusters are shown using lines connecting different characters (colour-coded to visualise connected components), the text-character associations are shown using dashed red lines and text-tail associations are shown using dashed purple lines. Images taken from [16].



Pre-Grounded Caption

The scene is set in what appears to be an urban alleyway with a brick wall in the background. (A character)[1] with medium-length hair is depicted sitting on the ground. (He)[1] is wearing a long-sleeve shirt and pants, and (his)[1] facial expression is contemplative. Next to (him)[1] is a dark round object, possibly a ball, and a small trash can. (Another character)[2] stands nearby, (his)[2] posture is relaxed with one hand resting on (his)[2] hip. (He)[2] has short hair and is dressed in a jacket and pants. The atmosphere suggests a casual or reflective moment between (the characters)[1,2].

Normal Caption (without using overlaid boxes)

The image depicts a manga-style scene featuring two characters in a dimly lit environment, likely a cell or small room with stone walls. The character on the left has medium-length hair and is wearing a loose outfit with long sleeves, sitting on the floor, leaning slightly forward, seemingly engaged in conversation. The character on the right is barefoot, with longer hair, and dressed in a simple, light-colored shirt and dark pants. This character is seated cross-legged next to a ball and chain, suggesting they might be imprisoned. In between them, a candle is lit, casting light and shadows around the room. The overall atmosphere is intimate and somewhat somber.

Figure 11. **Pre-Grounded Captions.** GPT-4o-2024-08-06 is provided with a panel image that has character bounding boxes overlaid. The model is then instructed to caption the image, while outputting any character-referring phrases in (phrase) [ID] format. Note that while the characters have been successfully grounded in the predicted caption (left), the quality of this caption is significantly inferior when compared to the normal caption (right) which is obtained without overlaying the boxes. Image: Belmondo by Ishioka Shouei.

C. More on PopCaptions

As discussed in the main paper, PopCaptions is a novel dataset of 3,374 captioned manga panels, drawn from 15 randomly selected chapters of the PopManga test set [30] spanning 694 manga pages. Additionally for each of these pages, we provide the overall ground truth story.

The captioning was performed as a two-step annotation process to ensure quality in panel descriptions and character identification. Initially, captions were generated using gpt-4o, offering a solid foundation of scene, character, and emotion descriptions, though these AI-generated captions often contained errors, embellishments, and occasional inconsistencies, especially regarding character actions and emotions.

To refine these captions, a team of four annotators reviewed and edited each one, correcting misinterpretations, removing inaccuracies, and enhancing clarity in character references. Annotators also selected character-referencing phrases and drew bounding boxes around corresponding characters in each panel, linking text descriptions to visual elements. This detailed manual process took approximately 3 minutes per panel, totaling over 10,000 minutes across 3,374 panels, ensuring a high-quality final dataset.

To obtain the page-level stories, we utilised the same pipeline proposed in this paper i.e. zero-shot prompting of an LLM, except all the information provided in the prompt was ground truth (annotated by humans) and therefore the LLM was only required to compile this information into a prose format. Specifically, we construct a prompt (detailed in Appendix D below) using human annotated captions, human annotated transcripts with ground truth character names, and instruct Deepseek-R1 [12] to generate the story. We then randomly sampled a subset of these for manual quality assurance and found the generated stories to be of high quality.

D. More on prompts

In this paper, we employed several prompts tailored to different tasks, such as prose generation (Sec. 5.2 of the main paper) and evaluation using language model judges (Sec. 7.3 of the main paper). These prompts were crucial for guiding the LLMs to produce outputs aligned with our objectives.

We do not claim that our prompts are optimal. The process of prompt design allows for a degree of artistic variation, and the vast space of possible prompts makes exhaustive ablation infeasible. Additionally, evaluating all variations is computationally and financially expensive. Despite this, we document the exact prompts we used to maintain transparency and reproducibility.

D.1. Prompt for prose generation

```
I have a series of manga panel descriptions and dialogues.

Panel 1

Description: {caption}
Dialogues: {transcript}

Panel 2

Description: {caption}
Dialogues: {transcript}

...

Panel N

Description: {caption}
Dialogues: {transcript}

I want you to write a summary so that a blind or visually impaired person can understand the story.
Make sure to stick to the provided details. All these panels belong to the same page so make sure
your narrative is coherent. The format of the narrative should be a prose.
```

D.2. Prompt for LLM judges

```
You are trying to tell if a candidate caption/prose is describing the same image as a reference caption/prose.
Given the following rubric, I want you to give me a score on a scale from 1-5.

### Rubric for Evaluating Manga Panel Descriptions (1-5 Scale)

1. **Severely Inaccurate (1):**
  - The predicted caption/prose is mostly unrelated to the reference. Key elements are either missing or
  incorrectly presented, and there are major contradictions that obscure the intended context.

2. **Somewhat Off-Base (2):**
  - The predicted caption/prose captures some correct ideas but overlooks many crucial details. Major inaccuracies
  exist, such as incorrect character features or setting descriptions. The overall theme may slightly resemble
  the reference but lacks precision.

3. **Partially Accurate (3):**
  - The predicted caption/prose includes several recognizable aspects of the reference but has important
  inaccuracies. While it conveys the general idea, significant details about multiple
  features like characters, actions, or settings may be misrepresented.

4. **Mostly Accurate (4):**
  - The predicted caption/prose captures most key elements of the reference accurately. Minor inaccuracies are
  permissible as long as they do not significantly alter the overall understanding. Additional thematic elements
  or details may be present if they enhance the scene without conflicting with its primary depiction.

5. **Highly Accurate (5):**
  - The predicted caption/prose is nearly identical to the reference in both content and context. It seamlessly
  captures every detail, and any enhancements serve only to enrich the description without deviating from
  the original meaning.

Predicted caption/prose:

{}

Reference caption/prose:

{}

Instructions:

Output your response in a json with a key "judgement" that contains your analysis of the predicted caption/prose
and a key "score" between 1 and 5 (decimal is allowed) that contains your score.
```

E. Can't industry-grade VLM models do it End-to-End?

An inquisitive reader might wonder: can't GPT-4o, Gemini etc. handle the entire task end-to-end? For instance, could we simply input a manga page and have the model generate a complete description of the story? This is a very reasonable question, and while the answer is "not quite", the possibility may not be far off in the future.

To illustrate, consider an example in Fig. 12, where GPT-4o-2024-08-06 is tasked with describing the story directly from a manga page. At first glance, the output may seem satisfactory; however, upon closer inspection, certain obvious issues emerge, as highlighted in the figure. Such challenges highlight the limitations of current models in understanding and conveying the full depth of manga storytelling in a single pass. Furthermore, the chosen example is relatively simple on purpose, featuring minimal action or chaotic elements. Instead, consider the example in Fig. 13, where the scene is more complex. We observe that these industrial-grade models do not currently possess the ability to re-identify characters across panels and struggle with narrative continuity.

This is where a step-by-step approach offers distinct advantages. By decomposing the problem, we can incorporate prior knowledge about manga, such as panel reading order or cultural conventions, as well as external resources like a predefined character bank. These inputs allow us to spoon-feed critical details to large models, reducing the likelihood of errors and enabling them to perform at their best. In this sense, we do not view ourselves as competitors to these models but rather as collaborators, integrating them into our pipeline to enhance their output.

Looking ahead, we can envision a future where such models are significantly improved by leveraging data generated from our approach. By using annotations mined through our pipeline, it may become possible to refine these models to better understand complex storytelling formats like manga. Thus, rather than striving for direct replacement, our method serves as a stepping stone toward more capable and reliable AI systems for manga understanding.

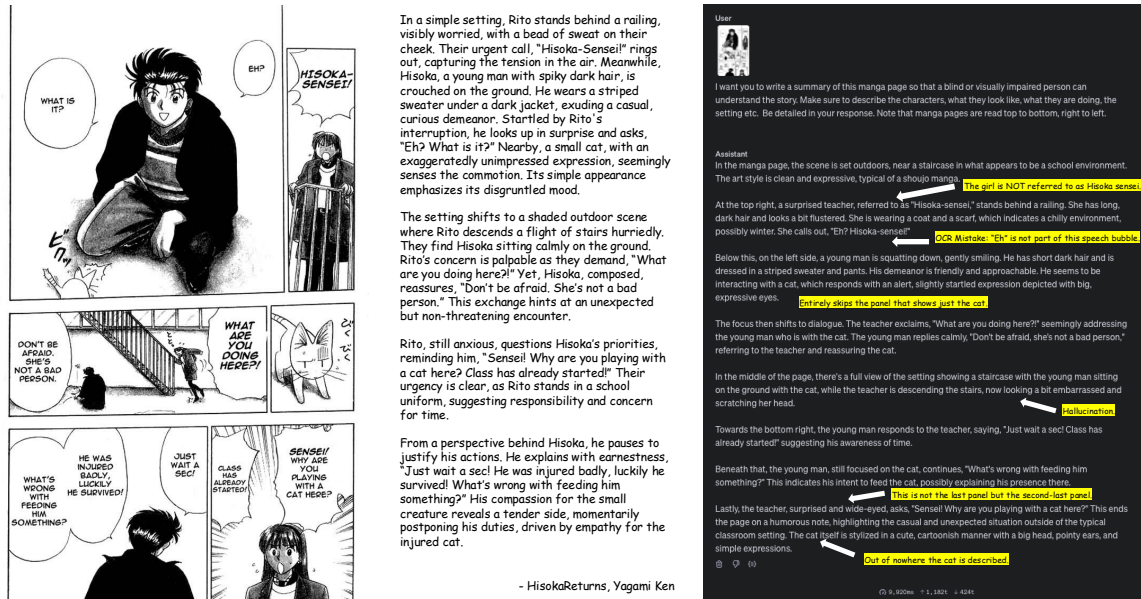


Figure 12. Comparison of our pipeline's output (left) vs directly using GPT-4o (right). Notice the issues highlighted in yellow.



Figure 13. Another example of End-to-End. Ignoring the style of the output, it is apparent that GPT-4o-2024-08-06 is not proficient at clustering characters across panels, and thus struggles with narrative continuity. Additionally, despite the instruction to follow top-bottom, right-to-left, reading order, the model fails to do so. Image: ©HanzaiKousyouninMinegishiEitarou, Kii Takashi.

F. Alternative narrative formats

While we chose prose as the primary format to describe manga for visually impaired readers, it is important to note that this choice is arbitrary. The flexibility of leveraging LLMs to generate the final output lies in their ability to adapt to alternative narrative formats, tailored to different audiences or purposes. For instance, as demonstrated in Fig. 14, we explore various narrative formats beyond prose. These include a children’s storybook format, which simplifies complex manga stories and makes them accessible to younger audiences. Such outputs can also incorporate age-appropriate content moderation, ensuring suitability for children. Additionally, we experiment with generating poetic descriptions, offering a creative and entertaining way to reimagine manga narratives.

Another compelling format is the script, designed to assist animators or film directors in adapting manga into video formats. Scripts can serve as a bridge between static visual storytelling and dynamic visual media, providing structured dialogue, scene descriptions, and action cues.

These examples highlight the versatility of LLMs in generating tailored narrative formats. By aligning the output format with the needs of specific audiences or applications, this approach demonstrates the potential of AI-driven storytelling to go beyond accessibility, offering innovative ways to reimagine manga for diverse contexts.

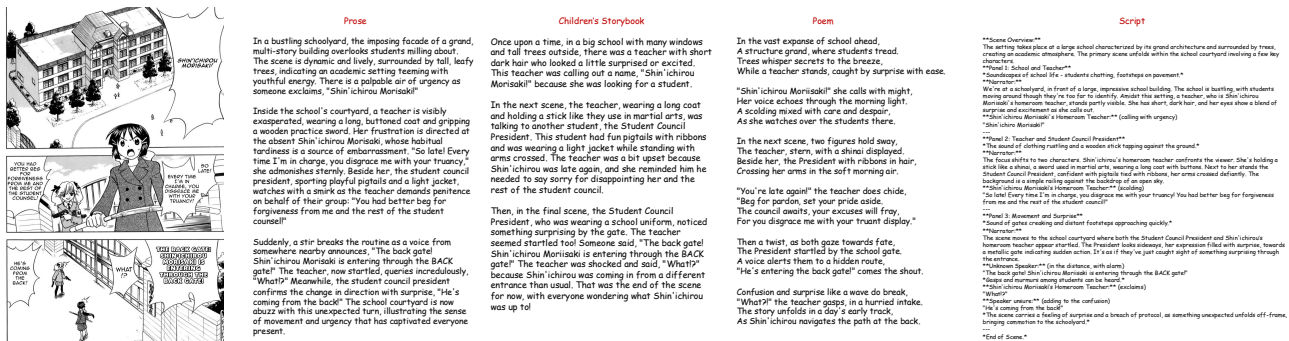


Figure 14. Illustration of various possible output formats. Image: ©Arisa, Yagami Ken.

G. Copyright and research on comics

As AI-driven research on comics becomes increasingly prevalent, it is crucial to consider the ethical implications, particularly regarding copyright and fair use. Copyright laws differ across jurisdictions, requiring researchers to carefully assess the legal frameworks governing the use of comic images in their respective countries. In our research, we utilise images scraped from the web, which is legally permitted for non-commercial academic research purposes in our country, aligning with practices used in large-scale datasets such as LAION and KITTI. However, to ensure ethical integrity and respect for creators’ rights, all images included in this manuscript are from Manga109 [2] where explicit permission is granted by the authors for use in publications. Researchers in this field should remain vigilant in navigating both legal and ethical considerations to promote responsible AI-driven scholarship on comics.

H. Examples of LLM scoring

In this section we provide some examples of reference prose, prose generated using our pipeline, and how the LLM judges critique the predictions.

Reference	Predicted	Scores
The scene unfolds with Light Yagami, a young man with sharp, focused eyes and medium-length hair, standing in tense contemplation. His collared shirt sleeve bears a distinct stripe, and a bead of sweat trails down his cheek as he raises a hand to his chin, deep in thought. The plain background amplifies the gravity of his expression[a mix of concentration and unease. In a close-up, his face tightens, eyes burning with intensity. The words escape him, raw and fractured; *I... I've killed two people...*. His voice hangs in the air, trembling with the weight of revelation. He steps back slightly, jacket draped over his shoulder, bag slipping as he clutches his chest. A dramatic shadow stretches behind him, mirroring his hunched posture like a specter. His speech splinters further: *I've... killed... Two people... ME...*. Each phrase fractures, as if the truth claws its way out against his will. Turning away, his profile reveals sweat glistening on his face, eyes wide with dread. The shadowed wall behind him feels oppressive as he whispers, *What do I do...*.a plea swallowed by the stillness. His hand suddenly grips a half-open bag, fingers digging into the fabric. The urgency in his voice sharpens: *Get... rid of this evil thing...*. The bag's contents remain hidden, but his desperation is palpable. The final moments plunge into chaos. Abstract shapes swarm the page[geometric fragments, jagged lines, and overlapping patterns evoking a fractured mind. In the void, Light's voice cracks, reduced to whispers: *\... No...*. The darkness swallows him, leaving only the echo of denial.	The page opens with Light Yagami, a young man with medium-length, textured hair, standing pensively. His hand rests at his mouth, eyes narrowed in deep thought, the lines of his jacket sharp against a sparse background. Though his expression is calm, faint strokes around his head hint at unease. No words break the silence here[only the weight of contemplation. The scene shifts to his profile, shadows deepening around him. His eyes widen slightly, intensity burning through their detailed pupils as he whispers, *I... I've killed two people...*. The admission hangs in the dark void behind him, stark and unresolved. Suddenly, we see him full-bodied, a bulky bag slung over his shoulder. His hands lift slightly, palms upturned in disbelief. A shadow on the wall mirrors his pose, stretching like a specter. His voice fractures: *I've... killed...*.a pause[*Two people...*.then, quieter, *ME...*. The building wall behind him feels cold, indifferent to his shock. In the next moment, his face reappears in profile, sweat beading his temple. His collar frames a tense jaw as he mutters, *What do I do...*. The background fades to white, isolating his anxiety. Action follows: a close-up of his hands fumbling with a zippered bag. Fabric wrinkles under his grip, sleeves rumpled. There's urgency here, his fingers trembling as he hisses, *Get... rid of this evil thing...*. The bag dominates the frame, a focal point of dread. The final panel dissolves into chaos[a jagged blur of shadows and jagged shapes. Screens or machinery loom abstractly, their forms swallowed by high contrast. Light's voice cuts through the darkness, sharp and final: *No...*. The page ends suspended in dissonance, his resolve and fear colliding.	Chatgpt: 4.5 Justification: The predicted prose and the reference prose both describe a scene featuring Light Yagami in deep contemplation, grappling with the revelation of having killed two people. Both descriptions focus on his intense expressions, the shadows and silhouettes around him, and his physical actions, such as holding a bag. The predicted prose accurately captures many critical elements like Light's emotional turmoil, the bag, and the abstract chaos at the end. However, there are some discrepancies and additional details in the predicted prose, such as the detailed description of the bag actions and the background's representation, which differ slightly from the reference. Despite these minor differences, the overall depiction and thematic elements align well with the reference. Gemini: 4.0 Justification: The predicted prose demonstrates a strong understanding of the reference, capturing most key elements and the overall progression of the scene. While there are minor deviations in specific details (e.g., the description of Light's clothing, the exact positioning of his hand, the nature of the abstract shapes in the final panel), these discrepancies do not significantly detract from the overall accuracy. The prose effectively conveys the emotional weight and thematic elements of the reference, showcasing Light's internal struggle and descent into chaos. The additions, such as the "faint strokes" around Light's head and the description of the wall as "cold, indifferent," enhance the scene without contradicting the original meaning. Therefore, the predicted prose deserves a score of 4, reflecting its high level of accuracy and insightful interpretation. Llama: 4.2 Justification: The predicted prose is mostly accurate, capturing key elements of the reference such as Light Yagami's emotions, actions, and setting. However, there are minor inaccuracies in details such as the description of his hair, jacket, and background. Gemma: 4.0 Justification: The predicted prose captures the essence of the panel description, hitting all major plot points and character emotions. The language used is evocative and mirrors the tone of the original. There are some minor variations in detail (e.g., description of Light's pose, the nature of the bag's contents), but the overall narrative arc and mood are faithfully conveyed.

Reference	Predicted	Scores
<p>The scene opens outdoors where a young boy named Son Goku stands frozen in surprise. His spiky hair ruffles slightly as he stares up at a massive, scaly pterodactyl-like creature looming over him. The beast's claws clutch a terrified young woman, Bulma, who screams in panic. The creature sneers, its voice booming with mock politeness: "Yeah...friend...that's it! Look, there's something I have to discuss with her in private, okay? You just wait right here!" Goku blinks, baffled, as the creature suddenly lunges forward, binding him tightly to a nearby tree with ropes. Struggling against the knots, Goku mutters, "Funny... WHY WOULD HE TIE ME UP?" His confusion lingers as the pterodactyl takes flight, its wings thrashing violently. Bulma writhes in its grip, her shouts muffled but desperate: "WHATA... VHA..." The creature cackles, "It's been far too long since I've tasted female human flesh!" Below, Goku remains tethered, his calm demeanor contrasting the chaos above. Bulma's distress crescendos. Tears stream down her face as she kicks and screams, "What are you WAITING for?! Rescue me, idiot!!!" Her voice pierces the air, urgency radiating from her every word. Meanwhile, the pterodactyl jeers, "DON'T TAKE TOO LONG, OKAY?" as it soars higher. Goku, still bound, tilts his head thoughtfully. The creature's taunts "imbecile! Dolt!" finally click. His eyes widen with realization. "Oh, so now she wants me to RESCUE her!" he exclaims, the urgency of Bulma's cries snapping him into focus. The page closes on this moment of clarity, tension hanging thick between Goku's restrained resolve and the fading laughter of the predatory beast.</p>	<p>On a forested outcrop, three figures collide in unexpected chaos. A winged pterodactyl-like creature, towering and sharp-toothed, coils its spiked tail around a young woman named Bulma, hoisting her off the ground. Her ponytail swings wildly as she shrieks in alarm, struggling to free herself. The creature cackles, its voice booming with false camaraderie: "Yeah...friend...that's it! Look, there's something I have to discuss with her in private, okay? You just wait right here!" Below them, a small boy with spiky hair[Son Goku]tilts his head, gripping a staff. His expression flickers between confusion and curiosity, unanswered questions hanging in the air. The scene shifts abruptly. Goku now sits bound to a tree trunk, ropes digging into his wrists. Despite his predicament, his face remains eerily calm. "Funny..." he murmurs, more puzzled than afraid. "Why would he tie me up?" The forest rustles around him, indifferent to his quiet bewilderment. Above the treetops, the pterodactyl soars, its jagged claws clamped around Bulma's dress. She dangles precariously, pigtailed whipping in the wind as the creature's laughter turns sinister. "It's been far too long since I've tasted female human flesh!" it roars, saliva dripping from its beak. Bulma's terrified screams ("WHA... VHA...") dissolve into the rushing air. Back on the ground, Bulma's patience snaps. In a burst of frustration, she stomps her feet, fists clenched and face flushed. "What are you WAITING for?!" she yells into the void, as if Goku could hear her across the distance. "Rescue me, idiot!!!" Her voice carries a mix of fury and desperation. Goku, finally untangling himself from the ropes, squints skyward. The pterodactyl's taunting laughter ("BWA-HA-HA-HA...") echoes as it dwindles to a speck. Goku scratches his head, the absurdity dawning on him. "Oh, so now she wants me to rescue her!" he mutters, half-amused, half-resigned. The forest hums with anticipation, leaves whispering as the boy grips his staff tighter, ready[or not]to leap into the fray.</p>	<p>Chatgpt: 4.5 Justification: The predicted prose closely resembles the reference prose by capturing most of the key elements, including the characters, the pterodactyl-like creature, and the essence of the scene. Both descriptions successfully convey the interaction between Goku, Bulma, and the creature, including the creature's deceptive friendliness, Bulma's distress, and Goku's bemusement leading to eventual realization. However, some minor discrepancies exist, such as the predicted prose adding thematic elements like Bulma's internal frustration and the forest's atmospheric reactions, which are not explicitly mentioned in the reference but enhance the narrative without altering its meaning. Overall, the predicted description maintains the core essence and details of the original scene while introducing additional context that complements the narrative.</p> <p>Gemini: 3 Justification: The predicted prose follows the general narrative flow of the reference but rearranges several key events and misrepresents some details. While it captures the core interaction between Goku, Bulma, and the pterodactyl, the sequence of actions is altered, impacting the portrayal of cause and effect. For example, the predicted prose has Goku tied up after Bulma is taken, while the reference clearly shows him being tied up first. Additionally, some dialogue is misattributed or missing, and the emotional nuances of the characters are not fully conveyed. The predicted prose introduces interpretations that are not explicitly present in the reference, like Goku's amusement at Bulma's predicament. While these additions don't necessarily contradict the scene, they add a layer of subjectivity that strays from the original depiction. Although the overall theme is preserved, the inaccuracies and reordered events prevent the predicted prose from achieving a higher level of accuracy. Thus, I am rating this a 3.</p> <p>Llama: 4.8 Justification: The predicted prose is highly accurate, capturing the key elements of the reference with precision. Minor inaccuracies exist, such as the order of events, but they do not significantly alter the overall understanding.</p> <p>Gemma: 3.5 Justification: The predicted prose captures the main elements of the reference: Goku encountering the pterodactyl, Bulma being kidnapped, and Goku eventually realizing he needs to rescue her. However, it deviates in some key areas. For example, it embellishes the scene with Goku being bound first, and Bulma reprimanding Goku for not rescuing her sooner. The overall plot progression is similar but has some added details that aren't in the reference.</p>

Reference	Predicted	Scores
In a dim school corridor late at night, Megumi Fushiguro pauses to catch his breath, his spiky hair damp with sweat and his high-collared coat rumpled. His eyes close briefly, shoulders slackening in weary relief after exertion. Nearby, Yuji Itadori watches him with a mix of curiosity and skepticism, his light spiky hair catching faint moonlight filtering through windows. As Megumi strides forward with renewed determination, motion lines ripple around his feet, betraying urgency. "Normally I'd be pissed," he admits gruffly, pausing to glance at Yuji, "...but good job." Yuji frowns, arms crossed. "Why're you all high and mighty?" he retorts, though his tone lacks malice. The tension shifts as two shadowy creatures materialize: one wolf-like with sharp ears devours a writhing mass of darkness, while another, paler beast gnashes its teeth nearby. Yuji blinks at the spectacle. "By the way, what's that eating the curse?" Megumi replies flatly, "It's my shikigami. You can see it?" Behind them, a small girl with glasses lies unconscious in Yuji's arms, her stillness contrasting the chaos. Megumi's gaze drops, his voice lowering as if lecturing himself. "Normally, you can't see Curses. Exceptions being when facing death or in special places like this." Yuji tilts his head, adjusting his grip on the girl. "Ohh... now that you mention it, I've never seen a ghost before," he muses, unfazed. A beat passes. Megumi studies Yuji's calm demeanor. "You're not scared, huh?" Yuji's expression softens, his focus drifting. "Well, I was..." He meets Megumi's eyes, voice steady. "Did you know? People actually die." Megumi stiffens, caught off guard by the weight of the statement. Around them, the corridor hums with residual energy the shikigami finish their task, the girl stirs faintly, and the two young men stand suspended between confrontation and camaraderie, their shared reality etched in shadows and unspoken questions.	The scene opens with Megumi Fushiguro, a young man with spiky dark hair and a high-collared uniform, exhaling in quiet relief. His eyes are closed, a faint sweat droplet on his cheek, as if he's just weathered something tense. Behind him, the room is sparse, with wooden floors and large windows framing Ryomen Sukuna, a figure with short, messy hair crouching nearby. Megumi walks away dismissively, muttering, \"Normally I'd be pissed... but good job.\" Sukuna retorts sharply, \"Why're you all high and mighty?\" A soft +fwoosh+ sound lingers in the air, hinting at unseen activity. The focus shifts to a hallway where Megumi stands calmly beside two wolf-like creatures one white, one dark sniffing at something invisible. Behind him, a light-haired person named Setsuko Sasaki cradles an unconscious figure in their arms, their expression curious. Another voice asks, \"By the way, what's that eating the curse?\" Megumi replies, \"It's my shikigami. You can see it?\" The question hangs unanswered. Megumi turns serious, explaining, \"Normally, you can't see Curses. Exceptions being when facing death or in special places like this.\" Sukuna, now wearing a hooded sweatshirt, looks startled, admitting, \"Ohh... now that you mention it, I've never seen a ghost before.\" His gaze drifts to Setsuko, who remains still, glasses askew, as if asleep. The tension shifts as Sukuna, leaning against a hallway wall, challenges Megumi: \"You're not scared, huh?\" Megumi faces him coolly, hands likely tucked in his dark coat. Sukuna adds, \"Well, I was... Did you know? People actually die.\" Megumi's brow furrows in a silent, wary \"Huh?\" the page closing on their charged standoff, the wolves and unseen forces lingering like shadows.	Chatgpt: 3.5 Justification: The predicted prose and the reference prose share several key elements but differ in specific details. Both describe a scene involving Megumi Fushiguro, a conversation with a secondary character, and the presence of wolf-like creatures identified as shikigami. However, there are notable differences: the secondary figure is identified as Ryomen Sukuna in the predicted prose and Yuji Itadori in the reference, which suggests a mix-up of characters. Additionally, the room and corridor settings vary, and the actions and responses of characters show some divergence in tone and context. The predicted prose introduces an additional character, Setsuko Sasaki, not present in the reference. These shifts indicate some inconsistencies with character identities, settings, and thematic undertones, but the general idea and tone are still partially captured. Gemini: 2.0 Justification: While the predicted prose captures the general storyline and some key dialogue, it suffers from significant character misidentification and inaccuracies in the setting and action details. Megumi's relief after exertion, the presence of Yuji (misidentified as Ryomen Sukuna and later Setsuko Sasaki), and the dynamic between them are core to the scene but are misrepresented. The description of Sukuna in a hooded sweatshirt is also incorrect. Although some thematic elements are preserved, like the discussion about curses and shikigami, the overall context is skewed by these errors. Llama: 4.2 Justification: The predicted prose is mostly accurate, capturing the key elements of the reference prose. However, there are some inaccuracies, such as the incorrect character name (Ryomen Sukuna instead of Yuji Itadori) and some minor details like the absence of motion lines around Megumi's feet. Additionally, the predicted prose omits some details like the moonlight filtering through the windows and the corridor humming with residual energy. Gemma: 2.0 Justification: The predicted prose significantly alters the characters and their interactions. Megumi and Sukuna are swapped for Megumi and Sukuna, and Setsuko Sasaki is introduced. Furthermore, the scene with Yuji and the unconscious girl is absent. While it captures the general mood, the specific details and character relationships are inaccurate.

Reference	Predicted	Scores
The night vibrates with primal energy as a colossal nine-tailed fox demon towers over a trembling forest. Kurama's massive form blots the moonlit sky, jagged teeth bared in a snarl that shakes the earth. Trees splinter beneath its paws, their trunks snapping like twigs as debris whirls in chaotic spirals around its thrashing tails. A guttural growl rumbles from its throat.*\Vy...\"a sound that seems to fracture the air itself. Below, human voices rise in desperation. A man shouts over the din, raw panic straining his words: *WE'VE GOT TO HOLD ON UNTIL LORD HOKAGE GETS HERE!*\". Bodies scramble[some lunging forward, others recoiling]as unseen forces toss a figure airborne. The ground trembles with every movement of the beast above, urgency sharpening the staccato breaths and clattering footsteps of those fighting to survive. Amid the chaos, Iruka Umino's voice cracks like a whip. He thrashes against arms restraining him, his muscles coiled with fury and fear. *\\et me go!!\"* he demands, the plea tearing from him as he strains toward an unseen battleground. *\\om AND DAD... THEY'RE STILL FIGHTING THE DEMON!!\"* His anguish hangs thick, a raw thread woven into the cacophony of crashing trees and distant roars. Suddenly, the perspective narrows[a single eye dominates the darkness, wide and unblinking. Shadows carve deep around its edges, the pupil sharp with resolve or dread. Every line of the face tenses, as if this silent witness teeters on the brink of a choice that could unravel the storm around them. The world holds its breath.	A towering nine-tailed fox demon snarls beneath a full moon, its massive form dwarfing the trembling forest below. Sharp teeth gleam as its nine swirling tails slice through the air like whips, radiating primal fury. Below this mythical threat, three shinobi brace themselves on a fractured battlefield. Two crouch low, their headbands glinting as they grit their teeth against an unseen force, while a third is hurled backward mid- air, robes flapping wildly. \"We've got to hold on until Lord Hokage gets here!\" one shouts, voice strained but determined. Nearby, a young boy with spiky hair thrashes against invisible restraints, his cloak billowing as he screams, \"Let me go!!\" Panic sharpens his features, but another voice cuts through the chaos[a companion clinging to him, desperate. \"Mom and Dad... they're still fighting the demon!!\" The plea hangs in the air, raw and urgent, as motion lines swirl around the boy's defiant stance. The scene shifts abruptly to a single, piercing eye framed by jagged stones. Its intensity burns with unspoken resolve[a silent vow or a moment of crystallized focus]amid the crumbling battlefield. The jagged rocks below the eye mirror the fractured world around them, shadows deepening as the fox's roar echoes across the moonlit night.	Chatgpt: 4.8 Justification: The predicted prose and the reference prose describe the same manga page scene with impressive detail and emotional intensity. Both descriptions capture the colossal presence of Kurama, the nine-tailed fox, and its impact on the environment, such as the trembling forest and the chaos created by its thrashing movements. Key elements such as the desperation of the shinobi, the critical shout about the arrival of Lord Hokage, and the emotional turmoil of Iruka Umino are similarly captured in both versions. While there are slight variations in wording, the thematic elements and actions described are highly consistent with each other. Additionally, both versions conclude with a shift to focusing on a singular eye, symbolizing resolve or tension against the backdrop of destruction. The enhancements and differences present do not obscure the primary depiction of the scene, and both accounts complement each other well. Gemini: 4.0 Justification: The predicted prose demonstrates a strong understanding of the reference image, capturing the central figures, their actions, and the overall atmosphere. It accurately depicts the nine-tailed fox demon, the struggling shinobi, and the distressed boy. The inclusion of dialogue and descriptions of the characters' emotional states enhances the narrative. However, it misses certain details, such as the specific name of the demon (Kurama), the detailed destruction caused by the demon, and the name of the boy (Iruka). The predicted prose also lacks the rich descriptions of sound and motion present in the reference. While it successfully conveys the scene's essence, the omission of these finer points prevents a perfect match. Llama: 4.2 Justification: The predicted prose captures most key elements of the reference accurately, but has some minor inaccuracies and creative liberties taken. The description of the nine-tailed fox demon, the shinobi, and the young boy's actions are mostly accurate, but some details like the setting, character features, and actions are slightly misrepresented or reimagined. Gemma: 4.0 Justification: The predicted prose captures most of the key elements of the reference: the nine-tailed fox demon, the scene of chaos and desperation, the presence of Iruka Umino, and the image of a resolute eye. It adds some details like the shinobi and their battle strategies, which can be considered additions to enhance the scene, but it slightly alters the focus by giving more prominence to the shinobi's actions.