Re:Verse - Can Your VLM Read a Manga?

Aaditya Baranwal†
University of Central Florida
Orlando, FL, USA
aaditya.baranwal@ucf.edu

Madhav Kataria
Indian Institute of Technology, Jodhpur
Karwar, RJ, IN
b23ch1025@iitj.ac.in

Naitik Agarwal
Indian Institute of Technology, Varanasi
Varanasi, UP, IN
naitikagrawal838@gmail.com

Yogesh Singh Rawat
University of Central Florida
Orlando, FL, USA
yogesh@crcv.ucf.edu

Shruti Vyas
University of Central Florida
Orlando, FL, USA
shruti@ucf.edu

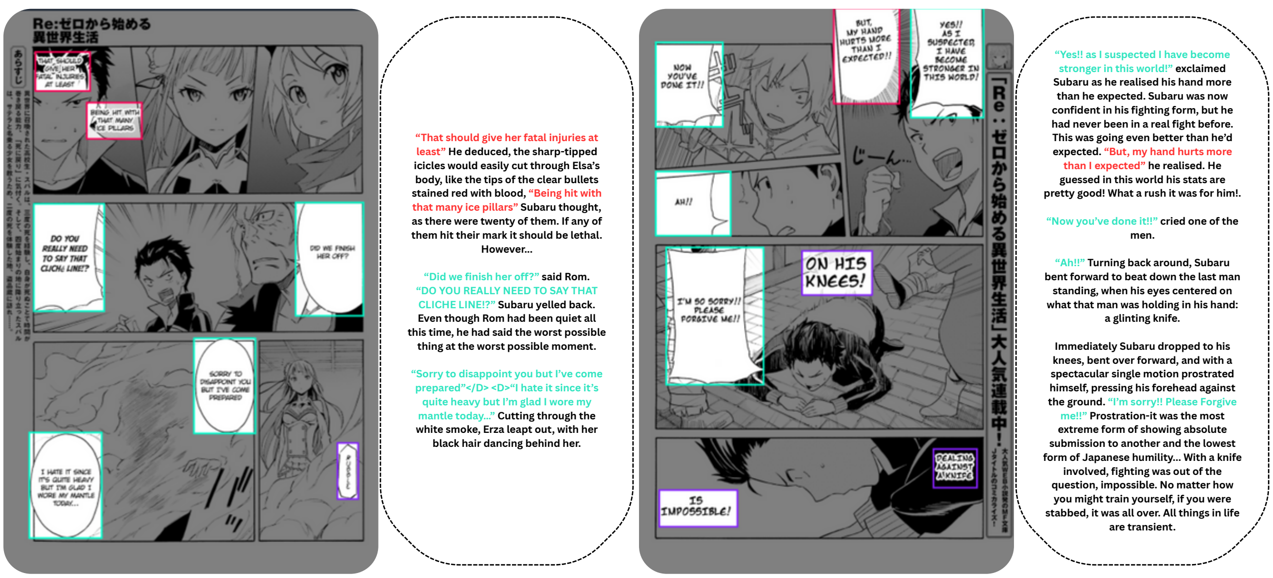


Figure 1. **Re:Verse Multimodal Annotation Examples.** Two representative examples from our dataset demonstrate the fine-grained alignment between visual manga content and narrative text. Each example shows the original manga page (left) paired with its corresponding aligned narrative text (right), where <D></D> tags indicate spoken dialogue (displayed in cyan) and <T></T> tags indicate internal thoughts (displayed in red). This precise cross-modal correspondence enables systematic evaluation of vision-language models' ability to understand sequential narrative structure, character consistency, and temporal progression. The semantic markup facilitates a comprehensive assessment of both surface-level text recognition and deep narrative comprehension capabilities across extended manga sequences.

Abstract

Current Vision Language Models (VLMs) demonstrate a critical gap between surface-level recognition and deep narrative reasoning when processing sequential visual storytelling. Through a comprehensive investigation of manga narrative understanding, we reveal that while recent large multimodal models excel at individual panel interpretation, they systematically fail at temporal causality and cross-panel cohesion, core requirements for coherent story comprehension. We introduce a novel evaluation framework that combines fine-grained multimodal annotation, cross-modal embedding analysis, and retrieval-augmented assessment to systematically characterize these limitations.

Our methodology includes (i) a rigorous annotation protocol linking visual elements to narrative structure through *aligned light novel text, (ii) comprehensive evaluation across multiple reasoning paradigms, including direct inference and retrieval-augmented generation, and (iii) cross-modal similarity analysis revealing fundamental misalignments in current VLMs' joint representations. Applying this framework to Re:Zero manga across 11 chapters with 308 annotated panels, we conduct the first systematic study of long-form narrative understanding in VLMs through three core evaluation axes: generative storytelling, contextual dialogue grounding, and temporal reasoning. Our findings demonstrate that current models lack genuine story-level intelligence, struggling particularly with non-linear narratives, character consistency, and causal inference across extended sequences. This work establishes both the foundation and practical methodology for evaluating narrative intelligence, while providing actionable insights into the capability of deep sequential understanding of Discrete Visual Narratives beyond basic recognition in Multimodal Models.*

Project Page: https://re-verse.vercel.app

Github: https://github.com/eternal-f1ame/Re-Verse

1. Introduction

Manga represents a rich and distinctive form of multimodal storytelling that blends expressive artwork, diverse panel layouts, and text embedded directly within images. Unlike standard natural images, understanding manga requires following visual and textual cues across multiple panels, interpreting characters' emotions, and tracking complex storylines [8]. These narratives often include non-linear timelines, inner thoughts, and symbolic visuals, making them uniquely challenging for computational models [36].

As vision-language models (VLMs) continue to advance multimodal understanding [16, 19, 23], applying them to manga opens exciting possibilities for supporting human creativity. Models that can understand manga could help creators edit, summarize, or expand their stories, acting as intelligent assistants that comprehend narrative flow [26]. To be truly useful, such models must follow context over many panels, recognize characters consistently, and reason about subtle visual and textual elements in a human-like manner. However, most existing benchmarks for manga understanding are limited in scope, focusing primarily on multiple-choice or classification-based tasks that operate at individual panel levels [1, 3, 12], without capturing broader narrative understanding or the full complexity of manga storytelling. This complexity often relies on events spread across multiple pages, emotional shifts, and implied information that requires sophisticated reasoning across extended sequences [9, 13]. To support the next generation of intelligent systems for manga, there is a critical need for benchmarks that evaluate models on their ability to follow long-range context, generate coherent summaries, and make narrative predictions for this intermediate modality of discrete visual narratives, sequences or exerpts. [31].

To address this fundamental gap, we introduce Re:Verse, a comprehensive benchmark for sequential narrative understanding in manga, designed to evaluate whether models can track and interpret story progression across entire chapters. Re:Verse is constructed from Arc 1 of Re:Zero [20, 21], a popular manga known for its intricate storytelling involving reincarnation loops and temporal resets. These narrative elements make Re:Zero an ideal testbed for evaluating long-range reasoning and panel-to-panel coherence in non-linear narratives [25]. Our benchmark contains 308 manually annotated panels from 11 chapters, with each panel including: (i) fine-grained bounding boxes and semantic tags for dialogue, thoughts, and key scene elements, and (ii) manually curated alignment with corresponding Re:Zero light novel passages, providing precise grounding (for detailed schematic procedure refer Figure 2).

We evaluate both open-source and proprietary models (with major focus on opensourced VLMs) [4, 22, 29, 39] across a comprehensive set of narrative tasks. Our findings reveal that while current models can extract surface-level semantics, they systematically struggle with maintaining narrative coherence, understanding character motivations, and handling temporal dependencies—particularly in stories with non-linear structures like Re:Zero [5, 33].

The contributions of this paper are:

- We introduce **Re:Verse**, a novel benchmark for manga narrative understanding, featuring 308 annotated panels with aligned visual and textual data that enables systematic evaluation of long-form narrative comprehension.
- We conduct comprehensive evaluation of state-of-the-art vision-language models across three challenging tasks: story generation, dialogue grounding, and temporal reasoning, revealing current fundamental limitations.
- We provide detailed analysis of model performance that characterizes key gaps in long-range coherence and multimodal narrative comprehension, establishing a foundation for future research in sequential visual storytelling.

2. Related Work

Computational analysis of comics and manga has evolved from structural detection to narrative comprehension, progressing through three stages: foundational datasets, task-specific benchmarks, and modern narrative understanding with Large Multimodal Models (LMMs).

Foundational Datasets for Structural Analysis

Early research established large-scale datasets for fundamental tasks. Manga109 remains influential, providing 109 volumes with detailed annotations for panels, text, characters, and faces [1], driving manga-specific structural analysis and computational comic understanding [8]. To diversify artistic styles, datasets like COMICS, eBDtheque [10], and DCM772 [32] introduced American and European

Table 1. Comparison of **Re:Verse** with existing benchmarks. Re:Verse provides chapter-length narrative evaluation with aligned textual grounding, enabling systematic assessment of story comprehension. *Task abbreviations:* NG = Narrative Grounding, DA = Dialogue Attribution, TR = Temporal Reasoning. *Evaluation types:* HE = Human Evaluation, AE = Automatic Evaluation, LLM-Based Evaluation.

Benchmark	Temporal Span	Tasks	Evaluation Focus	Aligned Text	Evaluation Methods
Manga109 [1]	Single	SP, OCR	Structural Analysis	–	AE
MangaUB [12]	Short-seq	VQA, PC	Panel Relations	–	AE (CircularEval)
MangaVQA [3]	2-Page	VQA	Spread Understanding	–	AE
CoMix [31]	Single	Multi-task	Integrated Reasoning	–	AE
Re:Verse (Ours)	**Chapter-length**	**NG, DA, TR**	**Narrative Intelligence**	✓	**HE, AE, LE**

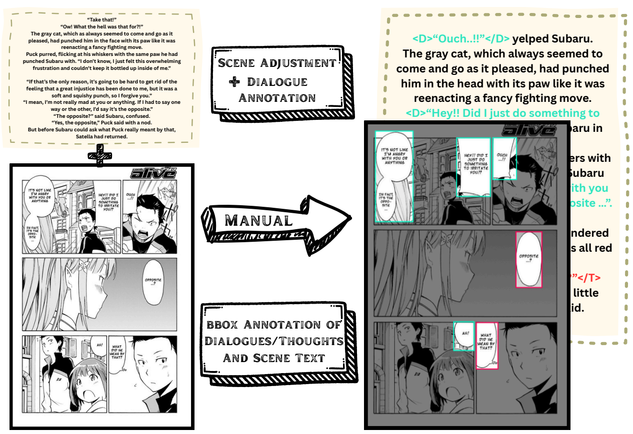


Figure 2. **Re:Verse curation and annotation pipeline.** Raw manga panels and light novel text are aligned via manual bounding box annotation, text classification (`<D>`, `<T>`), and semantic grounding, yielding synchronized visual, spatial, and narrative structure.

comics, though often with automated text extraction quality issues. COMICS Text+ addressed this with high-quality OCR benchmarks for Western comics [28]. These efforts established foundations for parsing visual grammar and understanding multimodal sequential storytelling [36].

Task-Specific Benchmarks and Integrated Reasoning

Subsequent work developed specialized benchmarks for integrated tasks. The Comic Onomatopoeia (COO) dataset extended Manga109 for recognizing stylized sound effects [2], while foundational work on "closure" between panels [13] established theoretical frameworks for sequential image comprehension. CoMix introduced comprehensive multi-task evaluation including speaker identification and dialogue generation [31], notably incorporating American comics to reduce manga over-representation and shifting focus from single-task to multi-task reasoning. Complementary efforts include M2C for manga complement generation [11] and MaRU for manga understanding [27].

Towards Narrative Comprehension with LMMs

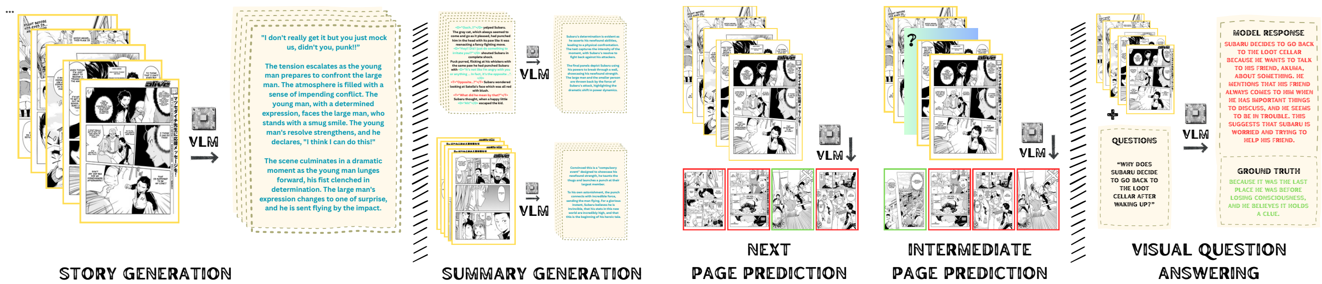
The advent of LMMs [4, 16, 19, 23, 39] catalyzed deeper narrative understanding evaluation. MangaUB performed fine-grained LMM analysis on single-panel recognition and multi-panel comprehension, identifying panel connections as key challenges [12]. MangaVQA advanced this with VQA tasks on two-page spreads, better emulating human reading and requiring cross-context synthesis [3]. Recent work explored sequential image-to-text reasoning [30] and cross-modal alignment [14, 17].

"From Panels to Prose" represents the most ambitious goal, generating full literary prose from manga pages including dialogue, descriptions, and actions [24]. Parallel developments in story evaluation [5, 33, 37] and narrative coherence assessment [6, 7, 34, 35] provided theoretical frameworks, while temporal reasoning benchmarks like TOMATO [25] addressed temporal causality challenges. The field benefits from cognitive science research on visual narrative structure [9] and comprehensive surveys of multimodal capabilities [15, 18, 38].

While these benchmarks have pushed the field from structural detection towards contextual reasoning, they predominantly operate on short sequences or isolated tasks (Table 1). A significant gap remains in evaluating the ability of LMMs to maintain **long-form narrative coherence** across entire chapters or story arcs. Our work, Re:Verse, is designed to fill this specific void. By providing a continuous, chapter-length dataset with aligned narrative text, it introduces the first benchmark focused on assessing temporal causality, character consistency, and story-level cohesion—aspects not systematically measured by prior work.

3. Re:Verse Benchmark

To address the fundamental limitations in existing benchmarks for sequential visual narratives, we introduce **Re:Verse**, a comprehensive benchmark designed to evaluate VLMs' ability to understand long-form manga narratives. Unlike existing benchmarks that focus on isolated panels or short sequences, Re:Verse systematically evaluates three critical aspects of narrative intelligence: story synthesis, character grounding, and temporal reasoning.



(a) Generative Setup: Story and Summary Generation.

(b) Temporal Reasoning: Page-Predictions and Visual Question-Answering.

Figure 3. **Overview of experimental setups used to evaluate narrative understanding.** (a) depicts the *Generative Tasks*, where a Vision-Language Model (VLM) is prompted to either (i) synthesize detailed narratives (**Story Generation**) or (ii) produce concise plot summaries (**Summary Generation**) (b) shows the *Temporal Reasoning Tasks*, (i) **Next Page Prediction**, where VLM selects the correct continuation of a story, (ii) **Intermediate Page Prediction** tests inference over missing narrative, and (iii) **VQA** for overall narrative understanding.

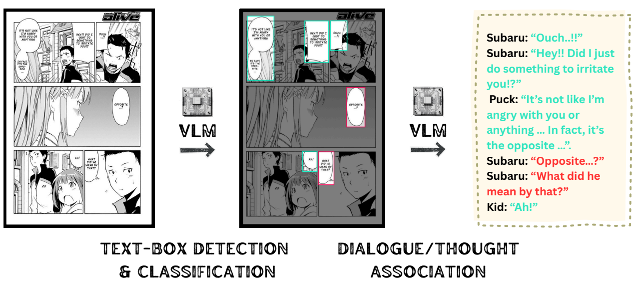


Figure 4. **Grounding Setup:** Text Detection, Classification, and Association with characters to evaluate in-page understanding.

3.1. Dataset Construction

We construct our benchmark from Volume 1 of *Re:Zero - Starting Life in Another World*, chosen for its complex narrative structure featuring time-loop mechanics, diverse character interactions, and sophisticated visual storytelling. The dataset comprises 308 meticulously annotated manga pages spanning 11 chapters, with each page containing:

Spatial Annotations. Every text bubble, thought bubble, and scene element is precisely localized with bounding boxes and semantic tags. This fine-grained spatial grounding enables evaluation of visual-textual integration.

Narrative Alignment. An expert annotator with comprehensive series knowledge manually adapted corresponding light novel passages to achieve one-to-one narrative correspondence with manga content. This alignment provides ground truth for story-level evaluation tasks.

Sequential Structure. Pages are organized to preserve narrative flow and temporal dependencies, enabling systematic evaluation of long-range coherence and causal reasoning across extended sequences as in the original work.

3.2. Analysis Framework

Our analysis framework systematically probes narrative understanding through three task categories (Figure 3):

Story Synthesis Tasks. We evaluate models' holistic narrative understanding through two generative tasks. *Story Generation* requires models to transform entire manga chapters into coherent prose narratives. *Cross-modal Summarization* evaluates models on summarizing both manga pages and their corresponding light novel text independently, quantifying comprehension and inference gaps and misses introduced in the visual storytelling modality.

Character Grounding Tasks. *Dialogue Extraction and Attribution* task requires models to spatially locate text elements, classify their semantic types (dialogue vs. thoughts), and correctly attribute them to characters using visual cues.

Temporal Reasoning Tasks. We frame temporal understanding through objective measures. *Next Page Prediction* (NPP) and *Intermediate Page Prediction* (IPP) evaluates models' ability to infer missing narrative events by predicting pages that bridge non-consecutive sequences. We predict the 6th and 11th page given the 5 and 10 consecutive prior pages, respectively, for NPP, and predict the missing page in a 2-missing-2 page and 3-missing-3 page setup for IPP. *Visual Question Answering* employs 55 complex questions requiring information across multiple pages.

3.3. Evaluation Methodology

Our evaluation protocol builds upon and extends prior frameworks such as SCORE [37] and AIStorySimilarity [7]. Unlike binary or pass/fail approaches, our methodology introduces proportional penalty systems, minimum score thresholds, and state-based incremental evaluation, enabling fine-grained differentiation of narrative quality. We employ both an LLM (GPT-4o [22]) and automated metrics (BERTScore, ROUGE, STTR, lexical density, NER density), with statistical significance testing and embedding-based cross-modal alignment (details in supplementary). Temporal reasoning tasks use accuracy measures within VQA frameworks. Cross-modal alignment is assessed through correlation analysis and embedding quality metrics using four vision encoders (BLIP, CLIP, SIGLIP, ALIGN).

Data Processing Pipeline. All manga pages undergo standardized preprocessing: resized to $224 \times width$ pixels while maintaining aspect ratio. Text annotations preserve dialogue tags $(<D></D>)$ and thought tags $(<T></T>)$ to maintain narrative structure distinctions crucial for evaluation. Character and entity names undergo manual validation using spaCy's named entity recognition enhanced with Re:Zero story-specific terminology (pronouns, salutations, honorifics, etc.).

Metric Standardization. NER density calculation uses character mention ratios to total word count. STTR employs sliding window approach (window size 50, step size 10) following standard text quality assessment. BERTScore utilizes bert-base-uncased with default settings for robust semantic similarity assessment. All results include confidence intervals ensuring statistical reliability.

This comprehensive evaluation framework enables systematic characterization of current VLM limitations in sequential visual narrative understanding, providing actionable insights for developing models with genuine narrative intelligence (understanding and generation) capabilities.

4. Analysis

We analyze VLM performance across story synthesis, visual-textual integration, and temporal reasoning. Our findings reveal systematic challenges in processing manga's *discrete temporal structure*, where narrative meaning emerges from discontinuous visual sequences. This analysis identifies the *inferent gap*—narrative discontinuity between pages requiring inference of missing story elements—as a fundamental bottleneck for current VLMs.

Model Selection Rationale. We focus on open-source VLMs rather than proprietary models for several critical reasons: (1) *Data contamination prevention*—Re:Zero's popularity likely exposes proprietary models to extensive training data contamination, making unbiased evaluation impossible; (2) *Reproducibility requirements*—academic benchmarks require stable, accessible models that enable reproducible research; (3) *Evaluation scale and cost*—comprehensive assessment across 308 pages with multiple tasks would be prohibitively expensive with proprietary APIs; and (4) *Democratization of comics understanding*—as open-source VLMs excel in video understanding, it is essential to democratize sequential visual narrative capabilities. This focus enables actionable insights for the broader research community while ensuring evaluation integrity.

4.1. Story Synthesis: Consistency Failures

Our story generation evaluation reveals systematic narrative coherence limitations across all models. Table 2 summarizes performance across six metrics, showing consistent gaps compared to human-authored narratives.

Table 2. Story generation performance across VLMs showing systematic character consistency failures and quality degradation compared to human narratives. NER density measures character mention frequency; STTR measures narrative tone and style.

Model	NER Density	STTR	Length	F1	ROUGE-1	Lexical Density
InternVL3-2B	0.023	0.71	89.4	0.803	0.143	0.54
Ovis2-2B	0.012	0.72	102.7	0.765	0.135	0.55
Qwen2.5-VL-3B	0.027	0.78	96.8	0.814	0.152	0.55
InternVL3-8B	0.012	0.81	95.1	0.817	0.151	0.56
Ovis2-8B	0.010	0.78	98.3	0.805	0.149	0.56
Qwen2.5-VL-7B	0.015	0.80	101.2	0.823	0.156	0.55
InternVL3-14B	0.009	0.81	97.6	0.819	0.165	0.57
Ovis2-16B	0.019	0.79	93.4	0.831	0.173	0.56
Gold Standard	0.087	0.82	156.2	1.000	1.000	0.61

Table 3. Character dialogue extraction and attribution performance demonstrating the critical disconnect between spatial text detection and character-dialogue grounding across VLMs.

Model	Recall	Precision	F1-Score	Character Accuracy
Qwen2.5-VL-3B	23.75%	21.84%	22.25%	1.11%
Ovis-8B	17.47%	9.89%	9.84%	0.85%
InternVL-14B	29.17%	20.65%	23.08%	0.00%

The most critical finding is systematic character consistency failure. Named Entity Recognition (NER) density ranges from 0.009 to 0.027 across VLMs, representing a 3-10x gap compared to the gold standard of 0.087. Even the best performing model (Qwen2.5-VL-3B) falls 69% short of human performance, indicating models cannot maintain character references across extended sequences. Quality metrics reveal additional degradation: Surface-form Type-Token Ratio (STTR) values range from 0.71 to 0.81 compared to 0.82 for human narratives, while generated content is systematically shorter (89-102 vs. 156 words) with lower lexical density. These patterns indicate models generate repetitive, formulaic content rather than nuanced narratives capturing rich character interactions.

Cross-modal Processing. Our cross-modal analysis (Table. 4 Summarization) compares corresponding content processing through text versus manga pages, revealing consistently close scores for automated metrics with gaps of just ranging from 1.1-3.2% BERTScore F1 points. Evidently, the close vicinity of the two scores establishes that automated metrics alone are nonsensical in the evaluation of narrative summarisation from visual content.

4.2. Character Grounding: Attribution Failures

We conducted a two-stage evaluation of dialogue extraction and character affiliation, revealing a critical breakdown in character-dialogue relationships. Table 3 shows volume-wide performance across all models, aggregated from chapter-level evaluations.

The results reveal a catastrophic disconnect between dialogue extraction and character understanding. While models achieve modest F1-scores for dialogue extraction (9.84%-23.08%), character attribution accuracy is near-zero

(0.00%-1.11%). Most strikingly, InternVL-14B completely fails to correctly attribute any dialogue to characters despite achieving the highest extraction F1-score of 23.08%, while Qwen2.5-VL-3B manages only 1.11% character accuracy.

These character attribution failures provide direct mechanistic evidence for the systematic character consistency gaps observed in our story generation experiments. The near-zero character accuracy scores demonstrate that current VLMs cannot establish the integrated visual-textual understanding necessary to ground dialogue in character identities—a fundamental requirement.

4.3. Visual-Textual Integration

We evaluate VLM performance on visual-textual integration and temporal reasoning tasks. Manga represents a unique *middle modality* where visual and textual information are spatially integrated, requiring simultaneous processing across multiple cognitive dimensions (Table 4).

Table 4 reveals critical patterns across all evaluation dimensions. For visual-textual integration, all models achieve reasonable spatial localization (IoU 0.580-0.630) but suffer severe text extraction failures with F1 scores below 0.25. This *visual-textual integration bottleneck* explains the visual processing penalties observed in summarization experiments. For temporal reasoning, InternVL3-8B shows remarkable improvement with longer context (43.2%→50.0% top-1 accuracy) while most models degrade, and the intuitive narrative constraint emerges where models consistently perform better on 3-missing-3 page scenarios than 2-missing-2 page ones.

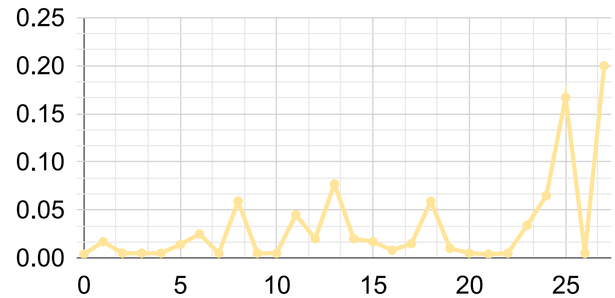
4.4. Temporal Reasoning: The Inferent Gap

Our temporal reasoning experiments directly probe the inferent gap through next-page and intermediate-page prediction tasks. The results reveal architecture-dependent processing differences and counterintuitive patterns that illuminate fundamental challenges in narrative understanding.

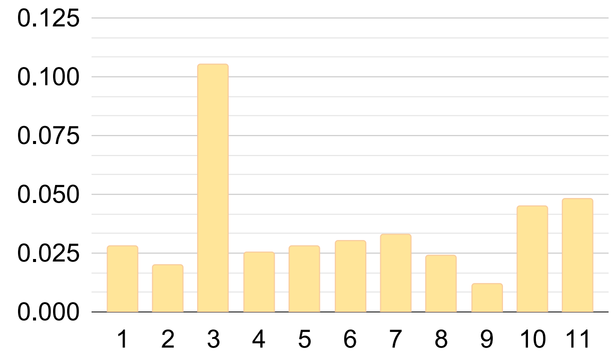
Architecture-Dependent Processing Differences. Our analysis reveals that architecture design impacts comics understanding more significantly than parameter count. The InternVL3-8B model's superior performance over larger models from other architectures suggests specialized temporal processing mechanisms are crucial. Notably, the Ovis2 series demonstrates dramatic scale-dependent improvement (16B model shows only 1.1 point cross-modal penalty vs. 2.2 points for 2B model), while Qwen2.5-VL maintains consistent 2.7-2.8 point penalties across scales, indicating architecture-specific scaling behaviors for visual narrative processing, even in non-performant scores.

4.5. Semantic Discontinuity in Narratives

Our retrieval experiments using four vision encoders (BLIP, CLIP, SIGLIP, ALIGN) provide quantitative evidence for



(a) **In-Chapter Analysis**



(b) **Over-Chapter Analysis**

Figure 5. **(a):** Mean Reciprocal Rank (MRR) across all chapters (x-asis: chapter number; y-axis: average MMR Scores), highlighting the cosistent overall poor scores on retrieving the visual counterpart provided the text. **(b):** Mean Reciprocal Rank (MRR) across pages of Chapter 7 (x-asis: page index; y-axis: MMR Score), highlighting the poor scores on retrieving the visual counterpart provided the text within the confines of a specific chapter.

semantic discontinuity in comics. Chapter-wise similarity analysis reveals highly variable semantic patterns demonstrating the inferent gap within and over the chapters.

Figure 6 shows that vision encoders struggle to maintain semantic continuity between consecutive pages. SIGLIP similarity scores range from -0.088 to 0.230, with negative scores indicating semantically orthogonal sequential pages. Volume-wide retrieval performance is consistently poor across all embedding methods (0.047-0.076 normalized similarity). CLIP achieves the highest performance at 0.076, representing less than 8% of theoretical maximum.

Chapter-Specific Embedding Patterns. Our analysis reveals systematic patterns in semantic discontinuity: early chapters (1-3) show highest semantic discontinuity with similarity ranges of 0.1-0.3 across all encoders, particularly challenging during character introduction sequences; mid-chapters (4-7) demonstrate moderate discontinuity (0.2-0.4 ranges) during plot development; while late chapters (8-11) achieve best coherence (0.3-0.5 ranges) in action-oriented

Table 4. **Comprehensive Performance across Visual-Textual Integration and Temporal Reasoning tasks.** Cross-modal penalties reveal processing degradation (TI = Text Input, VI = Visual Input), spatial localization succeeds but text extraction fails, while InternVL3-8B uniquely benefits from longer context and larger context gaps improve intermediate prediction.

	Summarization		Bubble Detection				Next-Page Pred.				Inter-Page Pred.			
Model	**TI**	**VI**	**IoU↑**	**Prec↑**	**Rec↑**	**F1↑**	**5p@1↑**	**10p@1↑**	**5pMRR↑**	**10pMRR↑**	**2p@1↑**	**3p@1↑**	**2pMRR↑**	**3pMRR↑**
InternVL3-2B	0.834	0.802	0.610	**0.394**	0.218	0.281	4.0%	8.9%	0.350	0.378	13.6%	18.2%	0.285	0.332
Ovis2-2B	0.833	0.811	0.580	0.065	0.054	0.059	40.9%	27.3%	0.581	0.477	22.7%	31.8%	0.398	0.477
Qwen2.5-VL-3B	0.837	0.810	**0.630**	0.088	0.051	0.064	20.0%	15.6%	0.479	0.411	15.9%	20.5%	0.318	0.367
InternVL3-8B	0.835	0.808	0.609	**0.472**	**0.255**	**0.331**	**43.2%**	**50.0%**	**0.716**	**0.750**	**36.4%**	**43.2%**	**0.523**	**0.614**
Ovis2-8B	0.844	**0.820**	0.608	0.057	0.049	0.053	35.5%	31.8%	0.531	0.477	25.0%	34.1%	0.432	0.523
Qwen2.5-VL-7B	**0.845**	0.817	**0.611**	0.221	0.263	0.240	25.0%	22.7%	**0.544**	**0.511**	18.2%	25.0%	0.356	0.432
InternVL3-14B	0.842	**0.823**	0.607	0.335	**0.352**	**0.343**	4.7%	13.6%	0.423	0.439	20.5%	27.3%	0.367	0.455
Ovis2-16B	**0.844**	**0.833**	0.609	0.234	0.231	0.233	**35.0%**	**29.5%**	0.567	0.523	**29.5%**	**36.4%**	**0.477**	**0.550**

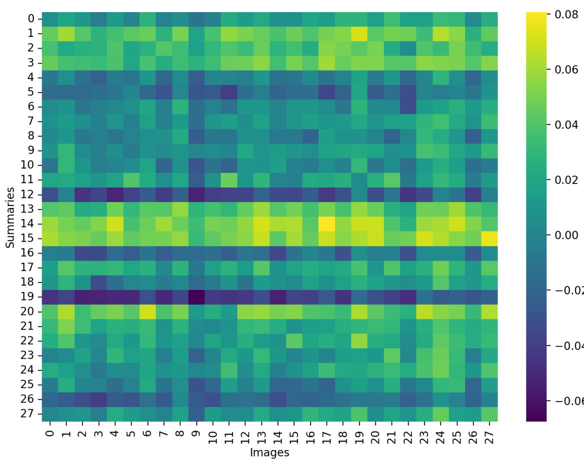


Figure 6. Chapter 7 similarity heatmap demonstrating semantic discontinuity patterns between manga pages and corresponding text using SIGLIP encoder. Irregular patterns reveal the inferent gap challenge in sequential narrative understanding.

sequences, suggesting current vision encoders are more amenable to visual action than narrative dialogue.

4.6. VQA: Temporal Reasoning Degradation

Our comprehensive Visual Question Answering evaluation across all 11 chapters provides direct evidence of systematic temporal reasoning failures (Table 5). The results demonstrate severe temporal reasoning limitations, with overall performance averaging only 1.43/5.0 (28.5% accuracy).

The results demonstrate severe temporal reasoning limitations, with overall performance averaging only 1.43/5.0 (28.5% accuracy). Critically, 7/9 models show performance degradation from early to late chapters (0.20-0.65 point drops). Chapter 11's narrative climax proves most challenging, with 8/9 models achieving lowest performance (0.80-1.20/5.0). This systematic temporal degradation provides direct evidence that VLMs struggle as narrative complexity accumulates across discrete visual sequences. High

Table 5. **VQA performance across narrative progression stages (0–5 scale).** *Later chapters reveal consistent degradation in reasoning ability. Bold-underline marks top 2 models overall.*

Model	**Ch 1-2↑**	**Ch 3-4↑**	**Ch 5-6↑**	**Ch 7-8↑**	**Ch 9-10↑**	**Ch 11↑**	**Overall↑**
InternVL3-2B	1.50	1.50	1.30	1.10	1.20	1.00	1.29
Ovis2-2B	2.00	1.50	1.30	1.40	1.00	1.20	1.42
Qwen2.5-VL-3B	1.80	1.30	1.60	1.00	1.20	1.20	1.36
InternVL3-8B	1.20	1.30	1.70	1.50	1.40	0.80	1.36
Ovis2-8B	1.80	1.60	1.40	1.40	1.50	0.80	1.47
Qwen2.5-VL-7B	**2.10**	1.20	1.30	**1.80**	**2.50**	1.00	**1.71**
InternVL3-14B	1.90	1.40	1.50	1.60	1.45	0.80	**1.50**
Ovis2-16B	1.80	1.50	1.50	1.50	1.20	0.80	1.44

chapter-wise variance (standard deviations 0.207-0.616) reveals inconsistent temporal understanding, indicating models process pages independently.

RAG Enhancement Analysis. To investigate whether external knowledge retrieval could mitigate these limitations, we evaluated models with and without retrieval-augmented generation (RAG). While RAG reduces complete failures (score 0) from 70% to 40% for models like Qwen2.5-VL-7B, high-quality responses (scores 4-5) remain extremely rare across all models. This demonstrates that temporal reasoning failures persist even with external knowledge augmentation, confirming that the core limitations stem from architectural inability to process discrete temporal structures rather than knowledge gaps.

These findings demonstrate that current VLMs cannot effectively process the discrete temporal modality defining comics as a unique narrative medium.

4.7. Novel Narrative Evaluation Methodology

To comprehensively assess VLM narrative generation quality, we introduce a **dual-framework evaluation system** building upon prior work (SCORE [37], AIStorySimilarity [7]) but with several key innovations. Unlike binary or pass/fail approaches, we introduce proportional penalty systems, minimum score thresholds, and state-based incremental evaluation, enabling fine-grained differentiation of narrative quality. We systematically evaluate four core narrative dimensions on a 0-100 scale: Character Development, Plot

Table 6. **Best narrative evaluation achieved by Ovis2-16B.** Higher values indicate better performance except for StdDev (↓).

Task	Overall↑	Character↑	Plot↑	Setting↑	Best↑	Worst↑	StdDev↓
Task 1	28.36	31.27	27.27	20.64	**31.25**	19.50	3.44
Task 2	28.34	**32.00**	**28.00**	19.09	**31.25**	**27.50**	**1.53**
Task 3	**28.70**	**32.00**	**28.00**	**21.64**	**31.25**	25.00	2.44
Average	**28.47**	**31.76**	**27.76**	**20.46**	**31.25**	**24.00**	**2.47**

Structure, Setting/Atmosphere, and Thematic Coherence. Score ranges provide clear quality differentiation from 90-100 (Exceptional) to 0-9 (Completely inadequate), enabling fine-grained comparison between models and revealing performance gradients invisible to binary evaluation systems. This represents the first penalty and score-based state metric for narrative quality assessment.

4.8. Narrative Coherence Breakdown

We conducted comprehensive narrative analysis using SCORE (Story Coherence and Retrieval Enhancement) framework and strict narrative similarity analysis. Table 6 presents OVIS2-16B evaluation results across all three tasks using our brutal narrative analyzer with penalties for narrative inconsistencies as parameterized in the framework.

Results reveal systematic narrative failures across all dimensions. With average overall score of 28.47/100, OVIS2-16B demonstrates fundamental limitations in constructing coherent long-form narratives. Character development (31.76/100) marginally outperforms plot construction (27.76/100), while setting development shows severe deficiencies (20.46/100). These findings confirm the *discrete temporal processing bottleneck* identified earlier. SCORE framework evaluation reveals VLMs process manga pages as isolated units rather than unified temporal narrative elements. Consistently low scores indicate current VLMs lack architectural components for maintaining narrative coherence across extended sequential contexts.

Narrative Depth Penalties. Our strict analyzer applies harsh penalties for insufficient story development. Observed scores reflect both poor narrative quality and fundamental inability to generate content with sufficient depth for meaningful story construction. This aligns with character attribution findings, suggesting VLMs struggle with hierarchical binding across multiple temporal contexts—a core manga understanding requirement.

5. Conclusion

We present Re:Verse, a comprehensive benchmark for evaluating VLMs on manga narrative understanding. Through systematic analysis across story synthesis, character grounding, and temporal reasoning, we demonstrate that current VLMs face fundamental challenges in processing the discrete temporal structure of visual narratives.

5.1. Key Contributions

Novel Benchmark Design. We introduce Re:Verse, the first benchmark for chapter-length manga narrative understanding. Unlike existing benchmarks focusing on isolated panels, Re:Verse features 308 annotated pages from 11 chapters with aligned narrative text, enabling systematic evaluation of long-form story comprehension.

Comprehensive Evaluation. We conduct extensive evaluation of 8 SOTA opensource VLMs across story synthesis, visual-textual integration, and temporal reasoning.

Identification of Fundamental Limitations. We identify three critical bottlenecks: (i) character consistency failures with 3-10x gaps in named entity recognition, (ii) visual-textual integration bottlenecks with F1 scores below 0.34, and (iii) temporal reasoning degradation with 28.5% accuracy on sequential understanding tasks.

Novel Evaluation Methodology. We introduce the first penalty and score-based state metric for narrative quality assessment, building upon SCORE [37] and AIStorySimilarity [7] frameworks. Our dual-framework system provides quantitative differentiation between poor-quality outputs rather than binary assessment.

Quantification of the Inferent Gap. First quantitative evidence for the *inferent gap* problem through intermediate-page prediction, showing models paradoxically perform better on 3-missing-page than 2-missing-page scenarios.

5.2. Impact and Future Directions

The systematic temporal reasoning degradation demonstrates models lack discrete temporal processing mechanisms necessary for narrative understanding. The Re:Verse benchmark provides a foundation for future research in sequential visual narrative understanding, extending beyond comics to other visual storytelling forms.

Broader Impact. This work represents a significant step towards enabling contextual understanding of manga and comics—culturally significant visual media that have been underserved by the vision community. Our benchmark enables development of context-aware applications for digital humanities research, creative assistance, and educational tools that may meaningfully engage with visual narratives.

** Re:Verse will be distributed under the CC BY-NC 4.0 license, permitting use for strictly non-commercial purposes. Access will be conditioned on explicit and mandatory acceptance of the license terms via a permission gate.*

6. Limitations

While Re:Verse establishes a new frontier, primary constraint is the scarcity of source material, given the high degree of narrative alignment between the *Re:Zero* manga and its light novel, the labor-intensive process required to create a high-fidelity dataset, ensures quality but limits its scale.

References

[1] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *IEEE MultiMedia*, 27(2):8–18, 2020. 2, 3

[2] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. Coo: Comic onomatopoeia dataset for recognizing arbitrary or truncated texts, 2022. 3

[3] Jeonghun Baek, Kazuki Egashira, Shota Onohara, Atsuyuki Miyai, Yuki Imajuku, Hikaru Ikuta, and Kiyoharu Aizawa. Mangavqa and mangalmm: A benchmark and specialized model for multimodal manga understanding, 2025. 2, 3

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3

[5] Christopher Cater and Mark Riedl. Fabula entropy indexing: Objective measures of story coherence. *arXiv preprint arXiv:2104.07472*, 2021. 2, 3

[6] Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation, 2022. 3

[7] Jon Chun. Aistorysimilarity: Quantifying story similarity using narrative for search, ip infringement, and guided creativity. In *Proceedings of the 28th Conference on Computational Natural Language Learning (CoNLL)*, pages 161–177, Miami, FL, USA, 2024. Association for Computational Linguistics. 3, 4, 7, 8

[8] Neil Cohn. *The visual language of comics: Introduction to the structure and cognition of sequential images*. Bloomsbury Publishing, 2013. 2

[9] Neil Cohn, Martin Paczynski, Ray Jackendoff, Phillip J Holcomb, and Gina R Kuperberg. (pea)nuts and bolts of visual narrative: Structure and meaning in sequential image comprehension. *Cognitive science*, 36(6):1084–1112, 2012. 2, 3

[10] Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1145–1149, 2013. 2

[11] Hongcheng Guo, Boyang Wang, Jiaqi Bai, Jiaheng Liu, Jian Yang, and Zhoujun Li. M2c: Towards automatic multimodal manga complement. *arXiv preprint arXiv:2310.17130*, 2023. 3

[12] Hikaru Ikuta, Leslie Wöhler, and Kiyoharu Aizawa. Mangaub: A manga understanding benchmark for large multimodal models, 2024. 2, 3

[13] Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daumé III, and Larry Davis. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1650–1659, 2017. 2, 3

[14] Byoungjip Kim, Sungik Choi, Dasol Hwang, Moontae Lee, and Honglak Lee. Transferring pre-trained multimodal representations with cross-modal similarity matching. In *Advances in Neural Information Processing Systems*, pages 13006–13018, 2022. 3

[15] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, and Michihiro Yasunaga. Vhelm: A holistic evaluation of vision language models. In *Advances in Neural Information Processing Systems*, 2024. 3

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022. 2, 3

[17] Jiaang Li, Yova Kementchedjhieva, Constanza Fierro, and Anders Søgaard. Do vision and language models share concepts? a vector space alignment study. *Transactions of the Association for Computational Linguistics*, 12:1232–1249, 2024. 3

[18] Jian Li, Weiheng Lu, Zhongzhi Xiong, and Xiaoyun Hao. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024. 3

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3

[20] Daichi Matsuse, Tappei Nagatsuki, and Shinichirou Otsuka. *Re:Zero - Starting Life in Another World, Chapter 1: A Day in the Capital, Vol. 1 (Manga)*. Yen Press, 2016. Illustrated by Daichi Matsuse. 2

[21] Tappei Nagatsuki and Shinichirou Otsuka. *Re:Zero - Starting Life in Another World, Vol. 1 (Light Novel)*. Yen Press, 2016. Translated by Jeremiah Borque. 2

[22] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng

Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. 2, 4

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual representations from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021. 2, 3

[24] Ragav Sachdeva and Andrew Zisserman. From panels to prose: Generating literary narratives from comics, 2025. 3

[25] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, 2025. 2, 3

[26] Conghao Tom Shen, Violet Yao, and Yixin Liu. Maru: A manga retrieval and understanding system connecting vision and language. *arXiv preprint arXiv:2311.02083*, 2023. 2

[27] Conghao Tom Shen, Violet Yao, and Yixin Liu. Maru: A manga retrieval and understanding system connecting vision and language, 2023. 3

[28] Gürkan Soykan, Deniz Yuret, and Tevfik Metin Sezgin. A comprehensive gold standard and benchmark for comics text detection and recognition, 2022. 3

[29] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[30] Danae Sánchez Villegas, Ingo Ziegler, and Desmond Elliott. Imagechain: Advancing sequential image-to-text reasoning in multimodal large language models. *arXiv preprint arXiv:2502.19419*, 2025. 3

[31] Emanuele Vivoli, Marco Bertini, and Dimosthenis Karatzas. Comix: A comprehensive benchmark for multi-task comic understanding, 2024. 2, 3

[32] Emanuele Vivoli, Irene Campaioli, Mariateresa Nardoni, Niccolò Biondi, Marco Bertini, and Dimosthenis Karatzas. Comics datasets framework: Mix of comics datasets for detection benchmarking, 2024. 2

[33] Heather Harris Wright, Gilson J Capilouto, and Anthony Koutsoftas. Evaluating measures of global coherence ability in stories in adults. *International journal of language & communication disorders*, 48(3):249–256, 2013. 2, 3

[34] Dingyi Yang and Qin Jin. What makes a good story and how can we measure it? a comprehensive survey of story evaluation, 2024. 3

[35] Jian Yao, Yiqun Cui, Dan Roth, and Yizhou Zhang. Openmeva: A benchmark for evaluating open-ended story generation metrics. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407. Association for Computational Linguistics, 2021. 3

[36] Victoria Yefymenko. Multimodality and cross-modal cohesion in manga. *Cognition, communication, discourse*, 24: 103–114, 2022. 2, 3

[37] Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Xinhang Yuan, Li Sun, Yi Xin, Jingqun Tang, Keqin Li, Kuan Lu, Menghao Huo, Jiaqi Chen, and Tianyu Shi. Score: Story coherence and retrieval enhancement for ai narratives, 2025. 3, 4, 7, 8

[38] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023. 3

[39] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2, 3