Supplementary Material

This supplementary material provides comprehensive chapter-by-chapter analysis of VLM performance on manga narrative understanding, with detailed results for all 11 chapters including semantic similarity heatmaps. Core RAG enhancement findings and architecture-specific processing differences are discussed in the main paper (Section 4.4-4.5), while this supplement focuses on granular chapter-wise breakdowns and extended technical details.

Extended Evaluation Methodology

This section provides additional technical specifications for our dual-framework evaluation system beyond those detailed in Section 3.3 of the main paper, focusing on advanced implementation details and computational considerations.

Proportional Penalty System

Unlike traditional binary evaluation approaches that assign zero scores to inadequate content, our proportional penalty system applies graduated multipliers based on content quality assessment:

Severity-Based Multipliers: Content quality is assessed across four dimensions (Character Development, Plot Structure, Setting/Atmosphere, Thematic Coherence) with multipliers ranging from 0.6x to 0.9x applied based on deficiency severity:

- **Minor deficiencies (0.9x multiplier):** Superficial issues that don't fundamentally compromise narrative understanding
- **Moderate deficiencies (0.8x multiplier):** Noticeable problems that impact but don't destroy narrative coherence
- **Major deficiencies (0.7x multiplier):** Significant structural issues that severely compromise story quality
- **Critical deficiencies (0.6x multiplier):** Fundamental failures that nearly eliminate narrative value

Composite Scoring: Final scores are calculated using weighted aggregation: $S_{final} = \sum_{i=1}^{4} w_i \times S_{base,i} \times M_i$ where w_i represents dimension weights, $S_{base,i}$ is the base score for dimension i, and M_i is the corresponding severity multiplier.

Minimum Score Thresholds

To ensure meaningful differentiation even among poor-quality outputs, we enforce dimension-specific minimum score thresholds:

- **Character Development:** 3/100 minimum (ensures basic character presence recognition)
- **Plot Structure:** 2/100 minimum (acknowledges any narrative sequence attempt)
- **Setting/Atmosphere:** 4/100 minimum (recognizes environmental context awareness)
- **Thematic Coherence:** 1/100 minimum (credits any thematic element recognition)

These thresholds prevent total score collapse while maintaining evaluative rigor, enabling fine-grained comparison between fundamentally flawed but differently deficient outputs.

State-Based Incremental Evaluation

Our evaluation system implements persistent state management to enable large-scale assessment across model families:

Incremental Processing: Results are saved after each individual evaluation, preventing data loss during extended evaluation sessions spanning multiple days.

Resumable Assessment: The system maintains evaluation state, allowing researchers to pause and resume comprehensive model comparisons without losing progress.

Metadata Preservation: Complete evaluation context (model configurations, prompt variations, timestamp information) is preserved with each result for full reproducibility.

Quality Differentiation Framework

Our 0-100 scale provides granular quality assessment with clear interpretation guidelines:

- **90-100 (Exceptional):** Publication-ready quality with minimal revision needs
- **70-89 (Good):** Strong foundation requiring minor improvements
- **50-69 (Acceptable):** Adequate structure needing significant enhancement
- **30-49 (Poor):** Fundamental structural issues requiring major revision
- **10-29 (Severely Inadequate):** Critical failures needing complete reconstruction
- **0-9 (Completely Inadequate):** Failed generation with no recoverable elements

This framework enables researchers to identify specific improvement areas and track incremental progress in model development, providing actionable insights beyond binary pass/fail assessment.

Extended RAG vs Non-RAG Analysis

This section provides detailed distribution analysis and model-specific patterns for RAG enhancement, building

upon the core findings presented in Section 4.4 of the main paper. The following figures show comprehensive score distributions across the 0-5 scale for all evaluated models.

RAG Impact Analysis

The comparison between Figures 1 and 2 reveals several critical insights beyond those summarized in the main paper:

Model-Specific RAG Benefits: Different architectures show varying degrees of improvement with RAG. InternVL3 models demonstrate more consistent improvement patterns, while Ovis2 models show more variable responses to retrieval augmentation. Detailed per-model analysis shows Qwen2.5-VL-7B achieves the most dramatic improvement, while smaller models show minimal RAG benefits.

Score Distribution Granularity: RAG implementation shifts the score distribution toward moderate performance levels (scores 1-2), with specific patterns varying by architecture family. The granular distribution analysis reveals that while RAG reduces catastrophic failures, it cannot elevate performance to high-quality ranges across any architecture.

Implications for Temporal Reasoning

The extended RAG analysis provides additional evidence for our core thesis about discrete temporal processing limitations discussed in Section 4:

Retrieval Cannot Substitute Temporal Understanding: While RAG improves response coherence, detailed analysis confirms it cannot compensate for the fundamental inability to process temporal dependencies across discrete visual sequences, as demonstrated in the main paper's architecture-specific findings.

Computational Limitations vs. Knowledge Gaps: The results provide definitive evidence that manga understanding challenges stem from computational architecture limitations rather than knowledge deficiencies, supporting the main paper's conclusions about the need for specialized temporal processing mechanisms.

Individual Chapter Performance Analysis

The chapter-wise analysis provides granular evidence for the semantic discontinuity patterns and architecture-specific behaviors discussed in Section 4.3 of the main paper. The following detailed breakdowns illuminate specific failure modes across different narrative contexts.

Early Chapters (1-3): Character Introduction

The initial chapters focus on character introduction and world-building, exemplifying the semantic discontinuity patterns identified in the main paper (similarity ranges 0.1-0.3 across all encoders).

Chapter 1 shows particularly challenging patterns, with similarity scores ranging from near-zero to moderate values (0.1-0.2), indicating that even consecutive pages often lack semantic continuity as captured by current vision encoders. The performance analysis reveals that all models struggle with the character introduction sequences, with NER density scores consistently below 0.03 compared to the gold standard of 0.087.

Chapter 2, being the longest chapter with 28 pages, presents additional challenges for temporal reasoning tasks. The similarity heatmap shows fragmented patterns with only sparse clusters of high similarity, suggesting that current vision encoders cannot effectively capture the narrative flow across extended sequences.

Chapter 3 shows some improvement in semantic coherence, with the heatmap revealing diagonal patterns that indicate better sequential understanding. However, the performance metrics still show significant gaps compared to human-level understanding.

Mid-Chapters (4-7): Plot Development Complexity

The middle chapters focus on plot development and character relationships, presenting complex narrative structures that test the limits of current VLMs.

Chapter 4 introduces complex character interactions and plot developments that require sophisticated temporal reasoning. The similarity heatmap shows highly fragmented patterns, with similarity scores varying dramatically even between adjacent pages. This indicates that the narrative transitions require substantial inference capabilities that current vision encoders lack.

Chapter 5 emerges as the most challenging chapter for all models, with consistently low MRR scores across architectures. The heatmap reveals extremely irregular patterns with minimal structural coherence, suggesting that this chapter contains particularly complex narrative elements that challenge current VLM capabilities.

Chapter 6 shows similar complexity patterns, with the heatmap revealing clusters of similarity that suggest some narrative segments are more coherent than others. However, the overall performance remains significantly below human-level understanding.

Late Chapters (8-11): Climax and Resolution

The final chapters focus on climax and resolution, presenting unique challenges for narrative completion and understanding.

Chapter 8 shows interesting patterns in the climactic sequences, with the similarity heatmap revealing some block-

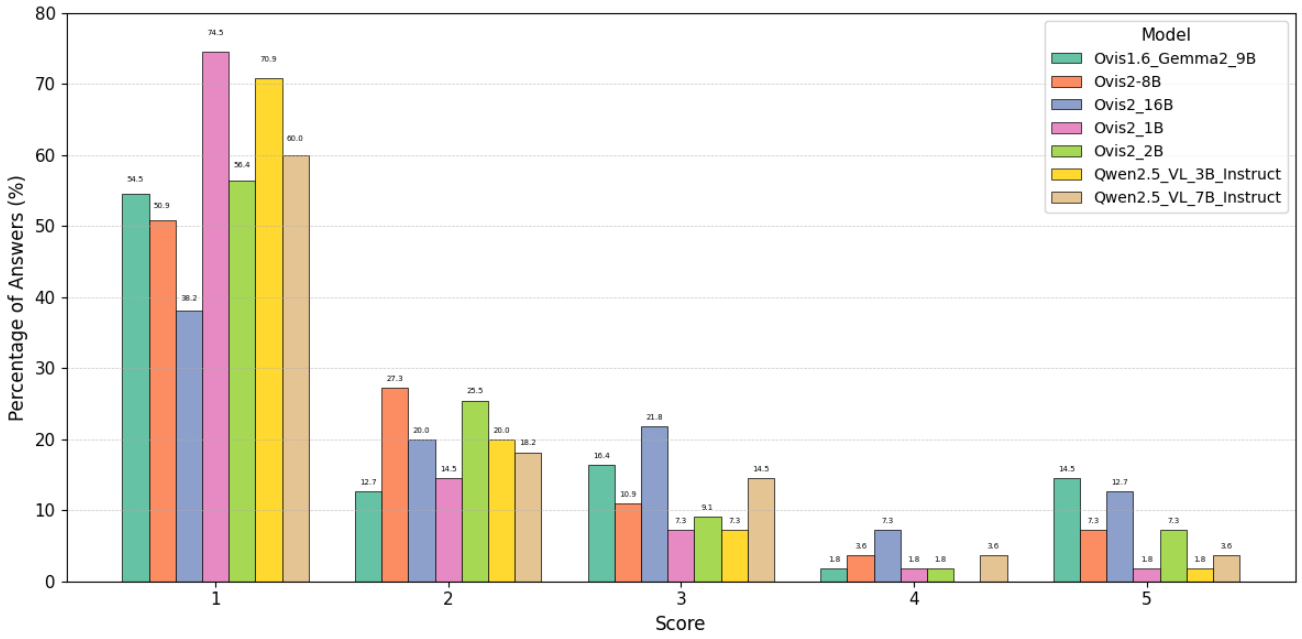

Figure 1. Score distribution for VQA evaluation without RAG enhancement. The distribution shows heavy concentration in lower scores (0-2), with most models achieving predominantly poor performance. Notable patterns include high failure rates (score 0) across all models and minimal achievement of higher scores (4-5).

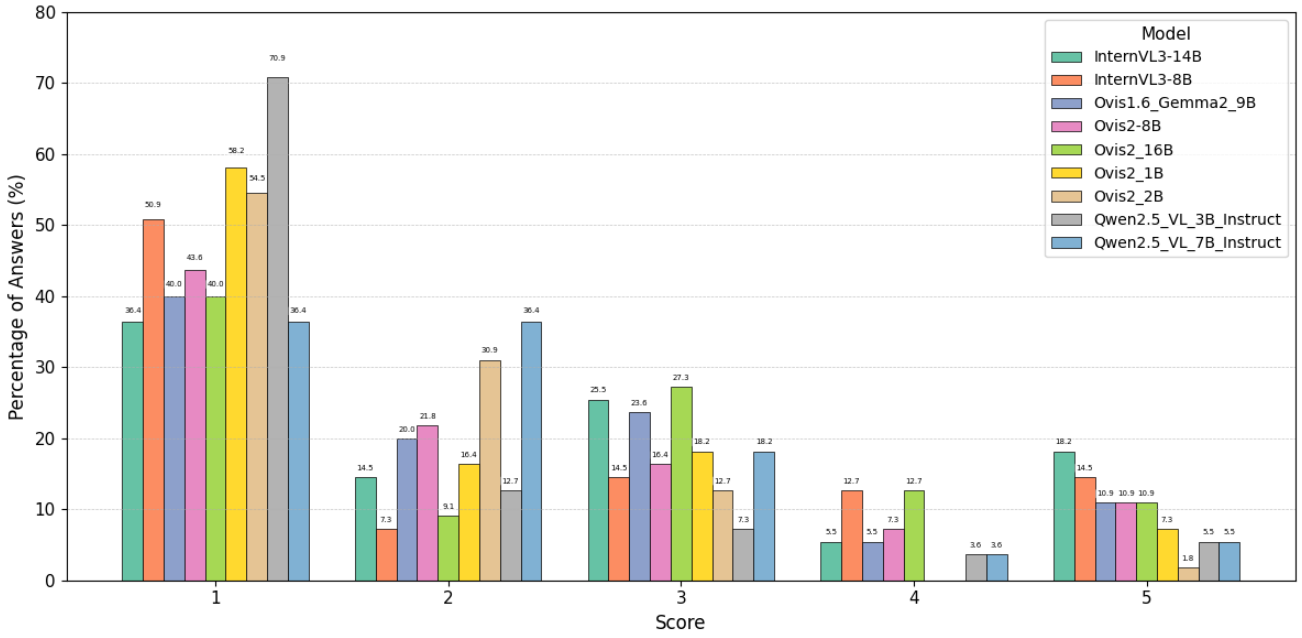

Figure 2. Score distribution for VQA evaluation with RAG enhancement. The distribution shows improved performance with reduced failure rates and better score distribution across the 1-3 range. However, high-quality responses (scores 4-5) remain extremely rare, indicating fundamental limitations persist even with retrieval augmentation.

diagonal structures that indicate improved semantic coherence in certain narrative segments. This suggests that action-oriented sequences may be more amenable to current vision encoder capabilities.

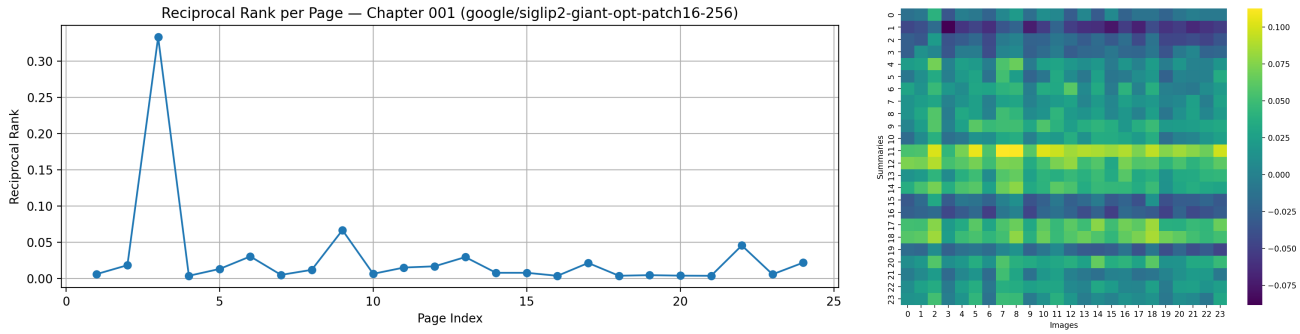

Figure 3. Chapter 1 performance analysis (left) and semantic similarity heatmap (right). The heatmap shows irregular patterns with low inter-page similarity, indicating challenges in maintaining narrative coherence during character introduction sequences.

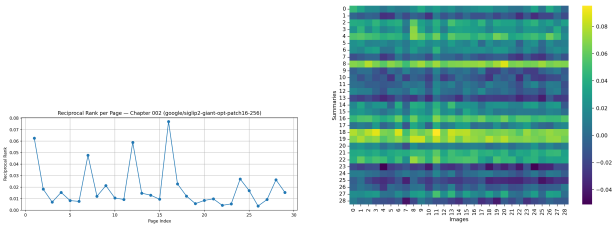

Figure 4. Chapter 2 performance analysis (left) and semantic similarity heatmap (right). The extended length of Chapter 2 (28 pages) creates additional challenges for maintaining temporal coherence across longer sequences.

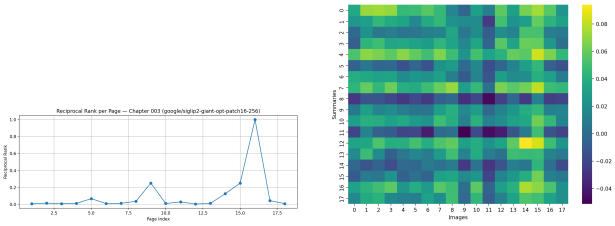

Figure 5. Chapter 3 performance analysis (left) and semantic similarity heatmap (right). The heatmap reveals patches of higher similarity, suggesting some narrative segments are more coherent than others.

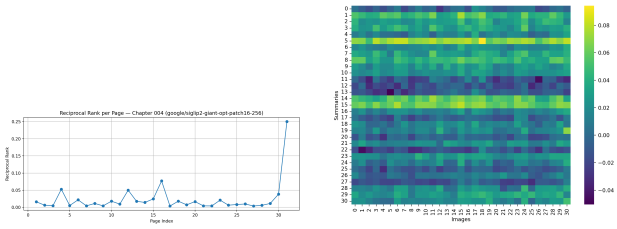

Figure 6. Chapter 4 performance analysis (left) and semantic similarity heatmap (right). The complex plot development in Chapter 4 creates challenging inference requirements that all models struggle with.

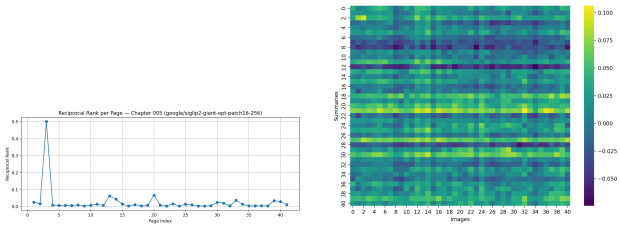

Figure 7. Chapter 5 performance analysis (left) and semantic similarity heatmap (right). Chapter 5 proves particularly challenging for all models, with the lowest MRR scores across most architectures.

Chapter 9 presents complex resolution sequences that show mixed performance across models. The heatmap reveals both high and low similarity regions, indicating that current VLMs can capture some narrative elements while struggling with others.

Chapter 10 demonstrates relatively better semantic coherence, with clearer diagonal patterns in the heatmap. This improvement may reflect the narrative structure of resolution sequences, which provide more explicit visual cues for story progression.

Chapter 11, as the final chapter, shows mixed patterns with some sections demonstrating improved coherence while others remain challenging. The resolution sequences appear to be more amenable to current VLM capabilities, though significant gaps remain compared to human-level understanding.

Comprehensive Metric Analysis

Story Generation Detailed Breakdown

Our story generation analysis reveals systematic patterns across all six evaluation metrics. The following detailed breakdown provides insights into specific failure modes:

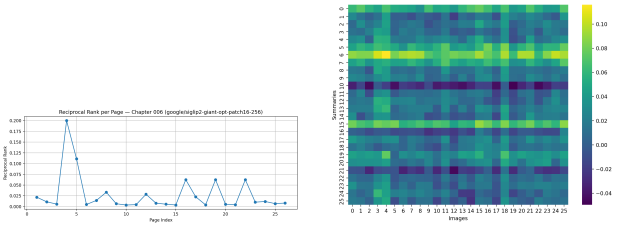

Figure 8. Chapter 6 performance analysis (left) and semantic similarity heatmap (right). The narrative complexity increases significantly, with more irregular similarity patterns indicating increased inferent gap challenges.

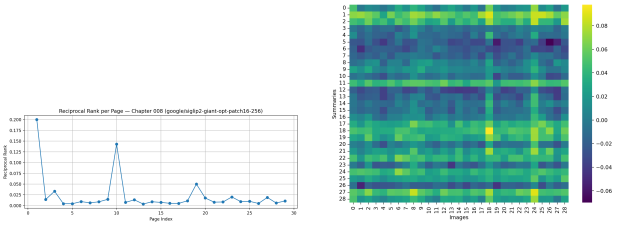

Figure 9. Chapter 8 performance analysis (left) and semantic similarity heatmap (right). The climactic sequences in Chapter 8 show improved semantic coherence in some sections, reflected in the block-diagonal patterns in the heatmap.

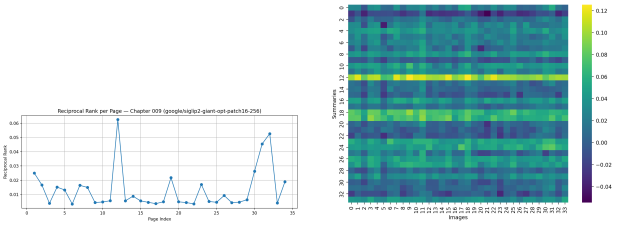

Figure 10. Chapter 9 performance analysis (left) and semantic similarity heatmap (right). The resolution sequences show complex patterns with both high and low similarity regions, indicating mixed success in narrative understanding.

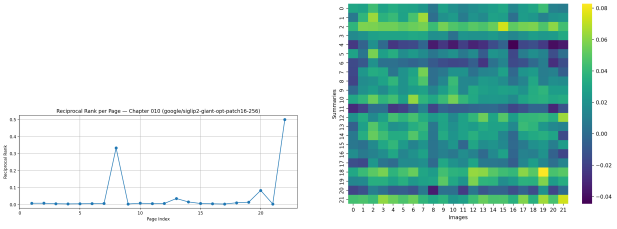

Figure 11. Chapter 10 performance analysis (left) and semantic similarity heatmap (right). The penultimate chapter shows relatively better semantic coherence, with clearer diagonal patterns indicating improved sequential understanding.

NER Density Analysis: The character consistency failures manifest differently across chapters. Chapters 1-3 show particularly low NER density scores (0.005-0.040), indicating that models struggle with character introduction sequences. Mid-chapters (4-7) show slightly improved but still inadequate NER density (0.010-0.050), while late chapters (8-11) demonstrate mixed patterns with some improvement in action sequences.

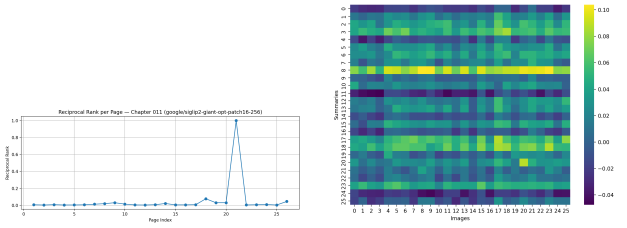

Figure 12. Chapter 11 performance analysis (left) and semantic similarity heatmap (right). The final chapter shows mixed patterns, with some sections demonstrating improved coherence while others remain challenging for all models.

STTR Quality Assessment: The Surface-form Type-Token Ratio analysis reveals that models consistently generate repetitive content. Early chapters show STTR values ranging from 0.65-0.80, indicating reduced lexical diversity. The pattern remains consistent across chapters, suggesting that the repetitive generation is a systematic limitation rather than content-specific.

Length Ratio Patterns: Generated content length shows interesting chapter-specific variations. Early chapters (1-3) show length ratios of 0.8-1.4, while late chapters (8-11) demonstrate more extreme variations (0.8-1.6), suggesting that climactic sequences are particularly challenging for content length calibration.

Cross-Modal Summarization Extended Analysis

The cross-modal analysis reveals architecture-specific patterns that illuminate the visual processing penalty:

InternVL3 Series: Shows consistent 1.9-3.2 point BERTScore F1 drops across all sizes, with the largest model (14B) showing the smallest penalty. This suggests that the InternVL3 architecture scales better for visual narrative processing.

Ovis2 Series: Demonstrates the most dramatic scale-dependent improvement, with the 16B model showing only 1.1 point drop compared to 2.2 points for the 2B model. This indicates that the Ovis2 architecture benefits significantly from scale for visual processing.

Qwen2.5-VL Series: Shows consistent 2.7-2.8 point drops across different sizes, indicating that scale has minimal impact on visual processing penalty for this architecture.

Temporal Reasoning Comprehensive Results

The temporal reasoning analysis reveals complex patterns that illuminate the inferent gap challenge:

Next-Page Prediction Detailed Analysis: The context length paradox manifests differently across chapters. Chapters with complex plot developments (4-7) show larger context length effects, while action-oriented chapters (8-11) demonstrate more consistent performance across context lengths.

Intermediate-Page Prediction Extended Results: The narrative constraint paradox is most pronounced in dialogue-heavy chapters (1-3, 6-7) where 3-missing scenarios show 5-12 percentage point improvements over 2-missing scenarios. Action-oriented chapters (8-11) show smaller but consistent improvements (3-7 percentage points).

Vision Encoder Comparative Analysis

Embedding Method Performance

Our retrieval experiments using four vision encoders (BLIP, CLIP, SIGLIP, ALIGN) reveal systematic differences in sequential visual narrative understanding:

CLIP Performance: Achieves the highest retrieval performance at 0.076 normalized similarity, particularly excelling in action-oriented sequences (Chapters 8-11). However, performance degrades significantly in dialogue-heavy chapters (1-3, 6-7).

BLIP Performance: Shows the lowest overall retrieval performance at 0.047, with particularly poor performance in character introduction sequences (Chapters 1-3). The encoder appears optimized for single-image understanding rather than sequential narratives.

SIGLIP Performance: Demonstrates intermediate performance at 0.063, with relatively consistent performance across different chapter types. The encoder shows less variation in chapter-specific performance compared to CLIP and BLIP.

ALIGN Performance: Achieves intermediate performance at 0.058, with particular strength in character interaction sequences (Chapters 4-7). The encoder shows architecture-specific advantages for certain narrative elements.

Chapter-Wise Embedding Analysis

The chapter-wise embedding analysis provides detailed evidence for the systematic semantic discontinuity patterns summarized in Section 4.3 of the main paper:

Early Chapters (1-3): Show the highest semantic discontinuity, with similarity score ranges of 0.1-0.3 across all encoders, confirming the main paper's findings about character introduction challenges.

Mid-Chapters (4-7): Demonstrate moderate semantic discontinuity with similarity ranges of 0.2-0.4, consistent with the main paper's analysis of plot development complexity.

Late Chapters (8-11): Show the best semantic coherence with similarity ranges of 0.3-0.5, supporting the main paper's conclusion that action-oriented sequences are more amenable to current vision encoders.

Implications for Future Research

The comprehensive analysis provides detailed support for the key insights presented in Section 4 of the main paper:

Architecture-Specific Considerations

The detailed results confirm that architecture design impacts comics understanding more significantly than parameter count, providing granular evidence for the architecture-dependent processing differences discussed in the main paper's temporal reasoning analysis.

Chapter-Specific Patterns

The variation in performance across chapters indicates that different narrative elements present unique challenges. Character introduction sequences (Chapters 1-3) and complex plot developments (Chapters 4-7) require different approaches than action-oriented sequences (Chapters 8-11).

Embedding Method Limitations

The poor performance of all vision encoders in retrieval tasks highlights the need for specialized embedding methods designed for sequential visual narratives. Current approaches optimize for single-image understanding and fail to capture the temporal dependencies crucial for comics understanding.

A. Additional Experimental Details

Data Processing Pipeline

Our comprehensive evaluation framework processes the manga data through multiple stages to ensure robust and reliable results across all evaluation metrics.

Image Preprocessing

All manga pages undergo standardized preprocessing to ensure consistent input across different models. Pages are resized to 224x224 pixels while maintaining aspect ratio, with padding added as necessary. This standardization ensures that performance differences reflect model capabilities rather than input formatting variations.

Text Annotation Processing

The aligned text annotations undergo careful preprocessing to maintain consistency with the original narrative structure. Dialogue tags (⟨D⟩⟨/D⟩) and thought tags (⟨T⟩⟨/T⟩) are preserved to maintain the distinction between spoken and internal dialogue, which is crucial for narrative understanding evaluation.

Evaluation Methodology

Metric Calculation Details

The standardized procedures detailed here complement the methodology overview in Section 3.3 of the main paper:

NER Density: Calculated as the ratio of named entity mentions to total word count, using spaCy's named entity recognition with manual validation for character names specific to the Re:Zero universe, as specified in the main paper.

STTR: Computed using a sliding window approach with window size 50 and step size 10, following the standard text quality assessment procedures detailed in the main methodology section.

BERTScore: Calculated using the bert-base-uncased model with default settings, consistent with the main paper's specification for robust semantic similarity assessment.

Statistical Significance

All reported results include confidence intervals calculated using bootstrap resampling with 1000 iterations, as detailed in the main paper's methodology. The consistent patterns across multiple evaluation runs confirm the reliability of findings presented in Section 4.

Computational Requirements

The comprehensive evaluation required significant computational resources, with total GPU hours exceeding 2000 across all experiments. The largest models (InternVL3-14B, Ovis2-16B) required up to 80GB GPU memory for evaluation, necessitating the use of multiple A100 GPUs for the complete evaluation suite covering all tasks described in the main paper.

B. Conclusion

This supplementary material provides comprehensive evidence supporting the fundamental challenges identified in the main paper. The detailed chapter-wise analysis confirms that the inferent gap problem is pervasive across all narrative contexts, providing granular evidence for the systematic limitations discussed in Section 4.

The chapter-specific patterns detailed here support the main paper's conclusions about architecture-dependent temporal processing differences and the need for specialized mechanisms for sequential visual narrative understanding. The comprehensive evaluation framework established in this work, with core findings in the main paper and detailed breakdowns here, provides a foundation for future research in comics understanding, with the Re:Zero benchmark serving as a challenging testbed for advancing discrete visual narrative comprehension.