Storyteller: Enhancing Interactive Storytelling via Real-Time Text and Image Generation without Additional Data

Chenrui Fan*
University of Bern
Bern, Switzerland
chenrui.fan@students.unibe.ch

Yi-Shiun Wu*
University of Neuchâtel
Neuchâtel, Switzerland
yi-shiun.wu@unine.ch

Lydia Y. Chen†
University of Neuchâtel / TU Delft
Neuchâtel, Switzerland
lydiachen@ieee.org

1 Introduction

The use of artificial intelligence in creative content generation has opened new possibilities for storytelling, art, and entertainment.[1],[2] This project aims to develop an innovative system that instantly generates a fully illustrated storybook based on minimal user input. By leveraging large language models (LLMs) for text generation and generative image models for visuals, the system transforms a short sentence provided by the user into a detailed narrative, accompanied by dynamically created illustrations.

The motivation behind this project stems from the need for an interactive, real-time storytelling experience in venues such as museums, libraries, and children's play areas. Traditional storybooks, while engaging, are inherently static and lack the ability to adapt to individual inputs or evolve based on real-time interactions. While pre-generating a collection of stories and matching them to user prompts could be a solution, it falls short in providing the personalized and creative experience that modern audiences seek. Pre-written content is limited in scope and may not align well with the diverse ideas users bring to the table. By contrast, real-time content generation allows each story to be uniquely tailored to the user's input, enhancing creativity, engagement, and immersion.[3]

The system we propose operates in three key steps: (1) the user inputs a brief sentence or phrase, (2) the LLM generates a coherent and engaging paragraph based on that input, and (3) an AI-based image model produces 3-5 illustrations that visually accompany the narrative. This approach fosters a personalized and dynamic storytelling experience, where each user interaction leads to the creation of a unique storybook, offering greater creative freedom and immersion compared to static story collections.

By integrating real-time text and image generation, the system enhances engagement in interactive venues, offering a unique, evolving experience where each interaction produces a one-of-a-kind story. This proposal outlines the technical approach, methodologies, and anticipated outcomes of building such a system, with a focus on providing a seamless and enjoyable user experience. The code for Storyteller is available at `https://github.com/Sosekie/Storyteller`

Figure 1: The existing results reveal issues. Three concatenated prompts, generated by GPT-4 based on a single sentence, are individually input into Stable Diffusion to produce corresponding images. However, the generated images are stylistically inconsistent and disparate. Each image requires 15 seconds to render and cannot be generated in parallel.

2 Methodology

2.1 Task Definition and Process Overview

The system generates a three-part story and corresponding images based on a single user-provided sentence.

2.2 Modalities and Data

Utilizing both text and image modalities, the system employs pre-trained Large Language Models (LLMs) and Latent Diffusion Models (LDMs) without additional training data, focusing on the effective use of existing models for high-quality outputs.

2.3 Input and Output Specifications

The system takes a sentence as input and produces three story segments, each with a corresponding image. The

*Equal contribution.
†Corresponding Author.

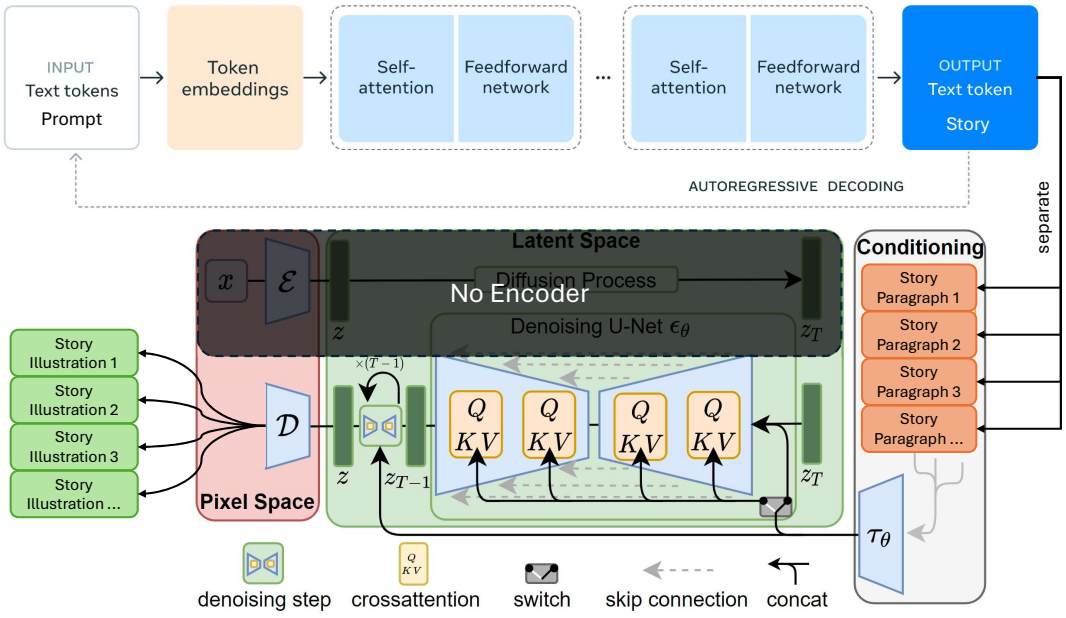

Figure 2: Structure.

process is as follows:

- The LLM generates a three-part story from the input prompt.
- The LDM generates an image for each story segment.

Which is shown in Figure 2.

2.4 Two Methods of Inference

A key challenge is the potential inconsistency in visual coherence among the three images, as characters or styles may differ significantly, which is shown in Figure 1. To address this, we considered two generation methods:

- **Queue-based Generation**: Each story segment is input sequentially into the LDM to generate the respective images. This often leads to significant differences in style and character appearance due to independent generation.
- **Many-to-Many Generation**: All three story segments are provided to the LDM simultaneously to generate the three images in parallel. A custom similarity metric is used to reduce inconsistencies between images.

2.5 Building the Latency/Throughput Model

To optimize the system for real-time generation, we build a latency and throughput model through the following steps:

- **Queueing System Simulation**: Implement a multi-worker queueing simulation to analyze average job latency and system performance under different dispatching strategies (Round Robin, Central Queue, and Join Shortest Queue). This approach evaluates the flow of tasks (LLM story generation and LDM image generation) and measures overall system efficiency in scenarios with varying load intensities and inter-arrival distributions.
- **Parameter Tuning**: Adjust system parameters (prompt type, prompt order, llm tokens, image size, generate type) to observe effects on job execution times. Use statistical methods like ANOVA to determine the significance of each parameter on latency and throughput.
- **Accuracy Verification**: Validate the queueing model's predictions against real-world data, refining assumptions as necessary to ensure the model accurately reflects system behavior.

2.6 Optimization for Real-Time Generation

To achieve real-time generation, we consider the following optimization strategies:

- **Parallel Inference**: Applicable to the many-to-many method, parallelization reduces latency.
- **Model Quantization**: Reduce model size via quantization to accelerate inference without significant quality loss.

2.7 Setup

All experiments were conducted on a single workstation equipped with an Intel Core i7-13700KF processor and an NVIDIA GeForce RTX 4090 GPU. The software environment included Python 3.10, CUDA 11.8, and the corresponding PyTorch framework, ensuring efficient computation and hardware-accelerated operations for both the large language model and latent diffusion model tasks.

3 Prompt Optimization

In our initial experiments, we observed that the narrative prompts generated by the Language Learning Model (LLM) were insufficiently detailed or contextually aligned for the Latent Diffusion Model (LDM), resulting in disjointed and thematically inconsistent illustrations, which are shown in Figure 3. Specifically, the second and third sentences of the generated story often replaced the main subject ("Mickey Mouse") with pronouns such as "he" or "his," causing the LDM to misinterpret the intended visual context. Additionally, certain prompt artifacts, such as repetitive or instructional statements, contributed to poor image coherence.

(a) Paragraph 1 (b) Paragraph 2 (c) Paragraph 3

Figure 3: Initial results without prompt cleaning and optimization. Subsequent figures demonstrate improved coherence after prompt refinement.

To address these issues, we introduced a prompt-cleaning stage before feeding the story sentences into the LDM. This stage removes meta-instructions (e.g., "You are a storyteller") and replaces pronouns with the proper noun "Mickey Mouse," thereby ensuring that each sentence explicitly references the main character. Such measures help maintain narrative continuity and improve the conceptual coherence of the generated illustrations.

Algorithm 1 shows the pseudo-code of our prompt cleaning and optimization approach. In essence, we remove repetitive or instructional cues and explicitly replace pronouns with the character name "Mickey Mouse." This ensures that every sentence directly references the main subject, providing the LDM with a clearer semantic grounding and leading to more contextually aligned illustrations.

3.1 Metrics for Quality and Coherence

In addition to prompt cleaning, we employ a comprehensive set of metrics to evaluate and ensure the quality, coherence, and performance of our system. By measuring both textual and visual properties, as well as timing information, we gain a holistic understanding of system behavior, as shown in Table 1.

By leveraging these metrics, we can refine our prompt generation and optimize the pipeline's overall performance. As shown in subsequent experiments, the enhanced prompts combined with systematic metric-driven evaluation lead to more thematically consistent stories and visually coherent illustrations, as illustrated by the sample outputs in Fig. 4.

Algorithm 1 Prompt Cleaning and Optimization

Require: Input prompt sentence I, number of sentences N, max tokens M

$P \leftarrow \text{GenerateBasePrompt}(I, N, M)$
$S \leftarrow \text{LLM}(P)$ ▷ Generate story text from LLM.
$R \leftarrow \text{SplitIntoSentences}(S)$
$U \leftarrow \{\text{"sentence"}, \text{"you need"}, \text{"prompt"}, \text{"story"}\}$
$R_{\text{filtered}} \leftarrow \emptyset$
for all $r \in R$ **do**
  if $\forall u \in U, u \notin r$ **then**
    $r \leftarrow \text{Replace}(r, \text{"he"}, \text{"Mickey Mouse"})$
    $r \leftarrow \text{Replace}(r, \text{"him"}, \text{"Mickey Mouse"})$
    $r \leftarrow \text{Replace}(r, \text{"his"}, \text{"Mickey Mouse's"})$
    Append r to R_{filtered}
  end if
end for
return $R_{\text{filtered}}[1:N]$

(a) Paragraph 1 (b) Paragraph 2 (c) Paragraph 3

Figure 4: Improved illustrations after prompt cleaning and metric-based optimization. Note the increased thematic alignment and overall narrative coherence.

4 Qualitative Results

This section showcases an illustrative example of the stories and images generated by the **Storyteller** system. The sample demonstrates the system's ability to produce coherent narratives and visually consistent illustrations based on minimal user input. By examining the example, we can assess the creativity, coherence, and visual quality of the generated content, providing a holistic view of the system's performance.

4.1 User Input

To demonstrate the functionality of our **Storyteller** system, consider the following user-provided sentence:

> *Mickey Mouse went to a western restaurant, ordered a roast chicken, and ate it.*

4.2 Generated Story

Based on the input, the Language Learning Model (LLM) generates the following three-sentence story:

> *Paragraph 1: Mickey Mouse went to a western restaurant, ordered a roast chicken, and ate it.*
>
> *Paragraph 2: Mickey Mouse walked into a small western town, where Mickey Mouse was greeted by a sign that read "Mickey's Diner".*

Table 1: Evaluation Metrics and Their Descriptions

Metric	Description	Goal
Time Metrics		
llm_time	Time taken by the LLM to generate the story.	Minimize
ldm_time	Time required by the LDM to produce images.	Minimize
pipeline_time	Total processing time from input to output.	Minimize
LLM Metrics		
vocab_diversity	Ratio of unique tokens to total tokens, indicating lexical richness.	Maximize
avg_sentence_length	Average number of tokens per sentence, reflecting sentence complexity.	Balanced
perplexity	Measures language model uncertainty; lower values indicate more fluent text.	Minimize
Story Context Metrics		
context_fid	Std. dev. of cosine similarities between consecutive sentences; lower indicates higher narrative coherence.	Minimize
avg_similarity	Mean cosine similarity between consecutive sentences; higher values suggest better thematic alignment.	Maximize
Image Quality Metrics		
avg_ssim	Structural Similarity Index Measure between images; higher values indicate greater image quality and similarity.	Maximize
diversity	Std. dev. of embedding distances among images; higher values indicate more varied and visually interesting results.	Balanced

Paragraph 3: Mickey Mouse ordered a roast chicken, and the waitress, a sassy cowgirl, handed it to Mickey Mouse with a wink.

4.3 Generated Illustrations

The Latent Diffusion Model (LDM) produced the following three illustrations to accompany the narrative:

(a) Paragraph 1 (b) Paragraph 2 (c) Paragraph 3

Figure 5: Sample Illustrations Generated by the Storyteller System

4.4 Analysis of Sample Output

The provided example illustrates both the strengths and areas for improvement of the **Storyteller** system:

- **Alignment with User Input**: The system successfully expands upon the user's prompt by situating Mickey Mouse in a specific setting—a western restaurant—and introducing engaging characters and scenarios. Despite the repetition, the story remains aligned with the original input, demonstrating the system's ability to create personalized and contextually relevant content.
- **Coherent Narrative Flow**: The three-sentence story maintains a logical progression, beginning with Mickey Mouse's visit to a western restaurant, his arrival in a small town, and his interaction with a sassy cowgirl waitress. This coherence aligns with the system's objective to generate engaging and structured narratives.
- **Visual Consistency and Quality**: The accompanying illustrations effectively represent key moments of the story—Mickey at the restaurant, the diner sign, and the interaction with the waitress. The images exhibit a consistent western theme, ensuring that the visual narrative complements the textual one seamlessly.
- **Engagement and Characterization**: The inclusion of a "sassy cowgirl" adds a dynamic character to the story, potentially increasing user engagement.

More output examples are provided in the Appendix.

5 Factorial analysis

5.1 Experimental Setup

We conducted a full-factorial experiment to systematically evaluate the effects of various parameters on the latency, textual quality, and image quality of the generated stories. Using the methodology described earlier, we designed an experiment that varied several key factors:

- **prompt_type**: (2 levels) e.g., *simple* vs. *detailed*.
- **prompt_order**: (2 levels) e.g., *input-first* vs. *prompt-first*.
- **llm_tokens**: (3 levels) The number of tokens requested from the LLM, e.g., {64, 128, 256}.

Table 2: ANOVA p-values for each dependent variable and factor. Significant p-values ($p < 0.05$) are in red.

Metric	prompt_type	prompt_order	llm_tokens	image_size	gene_type
Time Metrics					
llm_time	0.4151	1.84e-07	0.001257591	0.7220	0.6411
ldm_time	0.7830	0.9390	0.7505	5.34e-63	2.51e-77
pipeline_time	0.3090	0.02654382	2.52e-09	2.23e-30	1.18e-46
LLM Metrics					
vocab_diversity	0.046627	0.116221	0.723082	0.347718	0.911533
avg_sentence_length	0.367476	0.000235	0.251913	0.970289	0.302687
perplexity	0.8397912	8.70e-07	0.3931403	0.5579361	0.1237223
Story Context Metrics					
context_fid	0.140143	0.068635	0.671935	0.965527	0.363833
avg_similarity	0.960238	0.000174	0.407031	0.284691	0.886781
Image Metrics					
avg_ssim	0.6356562	0.4757334	0.7383935	1.82e-13	0.2190997
diversity	0.8088075	0.0831747	0.9655714	4.36e-11	0.2210438

- **image_size**: (3 levels) The resolution of generated images, e.g., {128, 256, 512}.
- **gene_type**: (2 levels) The image generation mode, e.g., *parallel* vs. *sequential*.

This gives a total combination of $2 \times 2 \times 3 \times 3 \times 2 = 72$ unique experimental configurations. To reduce randomness, we repeated each configuration three times and averaged the results.

Our input was a single user-provided sentence, from which the LLM generated a 3-sentence story. Subsequently, the LDM produced corresponding images. We recorded various metrics, which are also summarized in Table 1:

- **Time Metrics**:
  - *llm_time*: Time taken by the Language Learning Model (LLM) to generate the 3-sentence story.
  - *ldm_time*: Time taken by the Latent Diffusion Model (LDM) to generate the corresponding images.
  - *pipeline_time*: Total time from input to the final illustrated story output, encompassing both LLM and LDM processing times.
- **LLM Metrics** (computed from the generated story):
  - *vocab_diversity*: Ratio of unique tokens to total tokens, measuring the lexical richness of the generated story.
  - *avg_sentence_length*: Average number of tokens per sentence in the generated story, indicating sentence complexity.
  - *perplexity*: Approximation of the language model's uncertainty based on the distribution of token frequencies. Lower perplexity suggests more predictable and fluent text.
- **Story Context Metrics**:
  - *context_fid*: Standard deviation of cosine similarities between consecutive sentence embeddings, approximating the narrative coherence across the 3-sentence story. Higher values indicate more variability in coherence.
  - *avg_similarity*: Mean cosine similarity between consecutive sentence embeddings, measuring the thematic alignment and consistency between sentences. Higher values suggest better contextual alignment.
- **Image Quality Metrics** (computed from generated images):
  - *avg_ssim*: Average Structural Similarity Index Measure (SSIM) across all image pairs, assessing the structural quality and similarity of generated images. Higher SSIM values indicate greater similarity and quality.
  - *diversity*: Standard deviation of embedding distances among images, measuring the variation and richness of the generated image set. Higher diversity values signify a more varied set of images.

To analyze the results, we performed ANOVA tests for each metric to identify statistically significant factors.

5.2 Statistical Analysis

After running all configurations, we averaged repetition results and applied ANOVA. Table 2 provides a summary of significance levels (p-values) for each factor on selected metrics.

Here, we observe that:

- **Time Metrics**:
  - *llm_time* is significantly influenced by *prompt_order* and *llm_tokens*. Prompt-first ordering and higher token counts increase the LLM inference time.

- *ldm_time* is dominated by *image_size* and *gene_type*. Larger images and sequential generation drastically increase image generation time.
- *pipeline_time* is affected by multiple factors (*prompt_order*, *llm_tokens*, *image_size*, *gene_type*), indicating that the entire pipeline's runtime depends on both text and image generation parameters.

- **Textual Quality**:
  - *vocab_diversity* is slightly reduced by using a *simple* prompt type.
  - *avg_sentence_length* and *perplexity* are strongly affected by *prompt_order*. "Prompt-first" input leads to more complex, longer sentences and higher perplexity.

- **Contextual Metrics**: None of the tested factors strongly influenced *story_context_fid*, suggesting that context coherence is not easily controlled by these parameters. *avg_similarity* shows some sensitivity to *prompt_order*, indicating that how prompts are structured can weakly influence narrative consistency.

- **Image Quality**:
  - *avg_ssim* and *diversity* are significantly affected by *image_size* only. Larger images yield higher SSIM and greater diversity, suggesting image resolution is key for producing richer, more detailed visuals.
  - *gene_type* has a large impact on time but not on the final image quality metrics measured here.

5.3 Discussion

The results indicate that computational efficiency and content quality depend on different sets of parameters. For speed optimization, choosing smaller *image_size*, fewer *llm_tokens*, and parallel (*gene_type*) generation methods can yield faster turn-around times. Conversely, for richer textual and visual content, certain parameters stand out: *prompt_order* shapes the complexity of the text, and *image_size* strongly influences image richness.

While *prompt_type* slightly affects vocabulary diversity, its influence is limited compared to other factors. Similarly, though *gene_type* drastically changes the computational load, it does not significantly improve image quality metrics within the scope of these experiments.

In summary, the ANOVA analysis guides us toward strategic parameter choices for different objectives. If the goal is to minimize latency, one could reduce *llm_tokens*, use a prompt ordering that yields simpler text, and generate smaller images in parallel. If the objective is to maximize the narrative complexity or image detail, adjusting prompt order and increasing image size are more effective levers.

These findings offer valuable insights for fine-tuning the pipeline in a real-world storytelling setting, balancing performance constraints with narrative and visual fidelity.

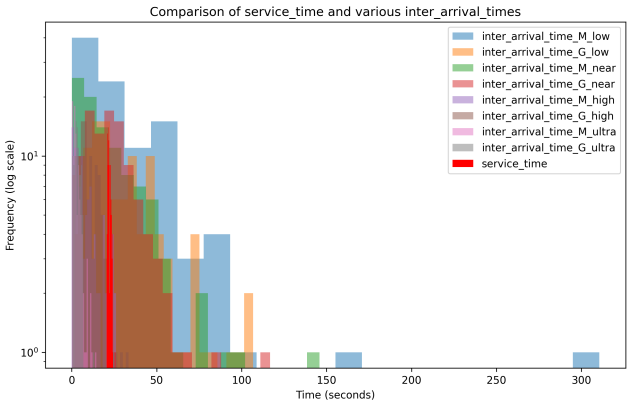

Figure 6: Histogram comparing the distributions of service times (red) and inter-arrival times (other colors) across various scenarios on a logarithmic frequency scale. Inter-arrival times include both exponential (M) and Gamma (G) distributions under four different load intensities: low, near saturation, high, and ultra-high.

6 Queueing Analysis

The primary objective of this section is to identify a queuing strategy capable of effectively handling varying levels of task load. Specifically, we aim to explore and evaluate the system's performance under diverse operational conditions, ranging from low utilization to near saturation and ultra-high load scenarios.

6.1 Experimental Setup

This study employs a queuing simulation framework designed to compare multiple queueing strategies under different arrival rates and inter-arrival time distributions. The dataset includes 100 samples, each providing service times and inter-arrival times for various scenarios. Details of the experimental setup are outlined as follows.

Service Time: Service times are drawn from a controlled distribution ranging from 20 to 26 seconds. To ensure internal consistency and minimize the influence of external factors, service times were generated using a fixed parameter pipeline:

- **Prompt Type:** Simple
- **Prompt Order:** Prompt-first
- **LLM Tokens:** 256
- **Image Size:** 512
- **Generation Type:** Parallel

This controlled approach ensures that variability in service times is primarily attributable to the system's intrinsic properties rather than extraneous influences.

Arrival Scenarios: To capture a broad range of operational conditions, four distinct arrival scenarios were defined:

Table 3: Theoretical Performance Metrics for Selected Scenarios

Scenario	λ	ρ	M/M/1 $E[T]$	M/M/1 $E[T_q]$	G/G/1 $E[W_q]$	Stability
M_low	0.0316	0.7005	74.02	51.86	39.25	Stable
G_low	0.0300	0.6657	66.31	44.14	10.85	Stable
M_near	0.0423	0.9383	359.30	337.13	169.43	Near-saturation
G_near	0.0392	0.8680	167.96	145.80	36.18	Near-saturation
M_high	0.1190	2.6391	-	-	-	Unstable
G_high	0.0996	2.2072	-	-	-	Unstable
M_ultra	0.3068	6.8018	-	-	-	Unstable
G_ultra	0.2938	6.5128	-	-	-	Unstable

- **Low Load:** Arrival rate of $1/30 \approx 0.0333$ tasks/s.
- **Near Saturation:** Arrival rate of $1/25 = 0.04$ tasks/s.
- **High Load:** Arrival rate of $1/10 = 0.1$ tasks/s.
- **Ultra-High Load:** Arrival rate of $3/10 = 0.3$ tasks/s.

Inter-Arrival Time Distributions: Each scenario was examined using two inter-arrival time distributions:

- **Exponential (M):** An exponential distribution with mean $1/\lambda$, where λ is the scenario-specific arrival rate.
- **Gamma (G):** A Gamma distribution with shape parameter 2.0 and scale parameter $(1/\lambda)$/shape, ensuring the mean matches $1/\lambda$.

The visualization of different arrival times is shown in Fig 6.

6.2 Statistical Analysis and Theoretical Performance

This subsection presents a statistical and theoretical analysis of queue performance under varying load conditions. We first compute key parameters, including mean arrival time, arrival rate (λ), service rate (μ), and utilization (ρ), and then apply M/M/1 and G/G/1 queueing models to estimate average response times ($E[T]$), average queueing times ($E[T_q]$ for M/M/1), and estimated waiting times ($E[W_q]$) for G/G/1 systems.

When the utilization nears or exceeds one (e.g., *inter_arrival_time_M_high*, *inter_arrival_time_G_high*, *inter_arrival_time_M_ultra*, *inter_arrival_time_G_ultra*), the system becomes unstable and theoretically unbounded waiting times are observed. Under such high-load conditions, a single-server M/M/1 or G/G/1 model does not effectively handle incoming tasks. Lower-load and near-saturation scenarios (e.g., *inter_arrival_time_M_low*, *inter_arrival_time_G_low*, *inter_arrival_time_M_near*, *inter_arrival_time_G_near*) exhibit increasingly degraded performance as utilization approaches one, but remain stable.

Table 3 summarizes the theoretical metrics for representative scenarios, highlighting the pronounced differences between low-load and high-load conditions and identifying the critical threshold at which standard queueing models fail to maintain stability.

Table 4: Average waiting times for different inter-arrival time scenarios.

Inter Arrival	Average Waiting Time (s)
M_low	19.235078
G_low	9.153372
M_near	73.531733
G_near	22.983422
M_high	664.039340
G_high	597.369679
M_ultra	946.841694
G_ultra	938.264642

6.3 Simulation Results and Analysis

Table 4 presents the average waiting times observed under different inter-arrival time distributions and load conditions. The results reveal a consistent pattern: as the system approaches saturation (i.e., utilization near or above one), waiting times increase sharply, eventually becoming exceedingly large when the arrival rate significantly exceeds the service capacity.

At low utilization levels (e.g., *M_low*, *G_low*), waiting times remain relatively short, with distributions that incorporate greater variability (G-type) yielding shorter queues compared to purely Markovian patterns. Near saturation (e.g., *M_near*, *G_near*), waiting times increase substantially, though G-type arrivals still exhibit a measurable advantage over M-type arrivals. Under high and ultra-high utilization scenarios (e.g., *M_high*, *G_high*, *M_ultra*, *G_ultra*), both arrival patterns produce excessively large waiting times, indicating that the system's single-server configuration cannot sustain the incoming load.

These findings highlight the importance of operating the system below critical utilization levels, where variability in inter-arrival patterns can effectively mitigate waiting times. They further suggest that once utilization surpasses a certain threshold, additional servers or more sophisticated scheduling policies become essential to maintaining acceptable queue performance.

6.4 Queuing Strategy

To mitigate prolonged waiting times under heavy workloads, we increased the number of workers from one to three and examined the impact of three distinct scheduling strategies:

Table 5: Average waiting times (in seconds) for different load scenarios and dispatching strategies.

Scenario	Round Robin (RR)	Central Queue (CQ)	Join Shortest Queue (JSQ)
M_low	**0.0463**	19.2351	**0.0463**
G_low	**0.0000**	9.1534	**0.0000**
M_near	**0.2703**	73.5317	**0.2703**
G_near	**0.0000**	22.9834	**0.0000**
M_high	**7.3104**	664.0393	**7.3104**
G_high	**1.1580**	597.3697	**1.1580**
M_ultra	**206.4217**	946.8417	**206.4217**
G_ultra	**196.9639**	938.2646	**196.9639**

1. **Round Robin (RR)**: Tasks are assigned CPU time quanta (26 seconds per task in this experiment). If a task completes before its quantum elapses, the CPU is immediately reclaimed; otherwise, once the quantum expires, control moves on to the next task. This cyclical approach ensures that no single job monopolizes the CPU for extended periods.

2. **Central Queue (CQ)**: All incoming jobs join a single central queue, and workers pull tasks from it as they become available. This approach relies on a first-come-first-served pattern without further load redistribution.

3. **Join Shortest Queue (JSQ)**: Incoming jobs are dispatched to the worker with the fewest pending tasks. This method dynamically balances the load by directing tasks towards less-busy workers.

Table 5 summarizes the average waiting times across various load conditions. Under low-load scenarios (*M_low, G_low*), both RR and JSQ achieve near-zero waiting times, markedly outperforming CQ, which incurs significant delays (e.g., approximately 19.24 seconds in *M_low*). As the load approaches saturation (*M_near, G_near*), RR and JSQ still maintain low waiting times (around 0.27 seconds or less), while CQ surges to 73.53 seconds (*M_near*), indicating a pronounced inefficiency when all tasks are funneled through a single queue.

Under high and ultra-high load (*M_high, G_high, M_ultra, G_ultra*), waiting times increase dramatically for all methods. Nonetheless, RR and JSQ remain substantially more effective than CQ. For instance, *M_high* scenarios see CQ reaching 664 seconds, compared to approximately 7.31 seconds for RR and JSQ. At even more extreme loads, RR and JSQ continue to provide considerable improvement over CQ.

Notably, due to the fixed 26-second time quantum and uniform distribution of tasks, RR and JSQ behave identically in these experiments. Under more heterogeneous conditions, JSQ's adaptive load balancing could potentially yield greater advantages. Overall, increasing the number of workers and employing distributed scheduling strategies (RR or JSQ) significantly reduces waiting times, especially as system load intensifies.

6.5 Summary of Queueing Analysis

This section investigated the performance of various queueing strategies under a wide spectrum of load intensities. By systematically adjusting arrival rates and employing different inter-arrival time distributions (Exponential and Gamma), we examined how simple single-server configurations become increasingly inefficient as utilization approaches or exceeds capacity. The analysis revealed that, while central-queue setups remain manageable at low loads, they incur substantial waiting delays as workloads intensify. In contrast, decentralized strategies—such as Round Robin (RR) and Join Shortest Queue (JSQ)—significantly improve system responsiveness. Under the conditions tested, RR and JSQ produced nearly identical outcomes, owing to fixed time quanta and uniform workloads. However, the flexible nature of JSQ suggests that it may yield further advantages in more heterogeneous scenarios. Ultimately, the findings highlight the critical importance of employing multiple workers and adaptive task allocation methods to maintain stable, low-latency performance across a broad range of operational conditions.

References

[1] Anantrasirichai, N., and Bull, D. Artificial intelligence in the creative industries: a review. *Artificial intelligence review 55*, 1 (2022), 589–656.

[2] Han, A., and Cai, Z. Design implications of generative ai systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (2023), pp. 470–474.

[3] Sun, Y., Xu, Y., Cheng, C., Li, Y., Lee, C. H., and Asadipour, A. Explore the future earth with wander 2.0: Ai chatbot driven by knowledge-base story generation and text-to-image model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), pp. 1–5.

Appendix

Additional output example I

Generated Story

"Paragraph 1: Mickey Mouse went to a western restaurant, ordered a roast chicken, and ate it.

Paragraph 2: As Mickey Mouse was finishing Mickey Mouse's meal, Pluto, the restaurant's loyal customer, approached Mickey Mouse.

Paragraph 3:"Mickey, I've been waiting for you," said Pluto, wagging Mickey Mouse's tail.

Generated Illustrations

The Latent Diffusion Model (LDM) produced the following three illustrations to accompany the narrative:

(a) Paragraph 1 (b) Paragraph 2 (c) Paragraph 3

Figure 7: Sample Illustrations Generated by the Storyteller System

Additional output example II

Generated Story

Based on the input, the Language Learning Model (LLM) generates the following three-sentence story:

"Paragraph 1: Mickey Mouse went to a western restaurant, ordered a roast chicken, and ate it.

Paragraph 2: As Mickey Mouse was finishing Mickey Mouse's meal, Pluto, the restaurant's dog, got up and started dancing.

Paragraph 3: Mickey laughed and said, "I guess that's one way to get a free meal!"

Generated Illustrations

The Latent Diffusion Model (LDM) produced the following three illustrations to accompany the narrative:

(a) Paragraph 1

(b) Paragraph 2

(c) Paragraph 3

Figure 8: Sample Illustrations Generated by the Storyteller System

Detailed Anova results

ANOVA results for llm_time:

	sum_sq	df	F	PR(>F)
C(prompt_type)	0.157603	1.0	0.672817	4.151156e-01
C(prompt_order)	8.011271	1.0	34.200725	1.839573e-07
C(llm_tokens)	3.479256	2.0	7.426604	1.257591e-03
C(image_size)	0.153401	2.0	0.327441	7.219664e-01
C(gene_type)	0.051389	1.0	0.219385	6.411004e-01
Residual	14.991535	64.0	NaN	NaN

OLS Regression Results

Dep. Variable:	llm_time	R-squared:	0.442
Model:	OLS	Adj. R-squared:	0.380
Method:	Least Squares	F-statistic:	7.229
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	2.27e-06
Time:	22:34:27	Log-Likelihood:	-45.673
No. Observations:	72	AIC:	107.3
Df Residuals:	64	BIC:	125.6
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.2712	0.161	7.880	0.000	0.949	1.594
C(prompt_type)[T.simple]	0.0936	0.114	0.820	0.415	-0.134	0.321
C(prompt_order)[T.prompt-first]	-0.6671	0.114	-5.848	0.000	-0.895	-0.439
C(llm_tokens)[T.128]	0.2045	0.140	1.463	0.148	-0.075	0.484
C(llm_tokens)[T.256]	0.5336	0.140	3.819	0.000	0.255	0.813
C(image_size)[T.256]	0.0598	0.140	0.428	0.670	-0.219	0.339
C(image_size)[T.512]	-0.0532	0.140	-0.380	0.705	-0.332	0.226
C(gene_type)[T.sequential]	-0.0534	0.114	-0.468	0.641	-0.281	0.174

Omnibus:	12.646	Durbin-Watson:	2.234
Prob(Omnibus):	0.002	Jarque-Bera (JB):	13.282
Skew:	0.977	Prob(JB):	0.00131
Kurtosis:	3.779	Cond. No.	5.89

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

ANOVA results for ldm_time:

	sum_sq	df	F	PR(>F)
C(prompt_type)	0.013484	1.0	0.076486	7.830089e-01
C(prompt_order)	0.001042	1.0	0.005911	9.389548e-01
C(llm_tokens)	0.101677	2.0	0.288369	7.504540e-01
C(image_size)	985.144617	2.0	2793.983700	5.337900e-63
C(gene_type)	2588.133492	1.0	14680.489880	2.510017e-77
Residual	11.283039	64.0	NaN	NaN

OLS Regression Results

Dep. Variable:	ldm_time	R-squared:	0.997
Model:	OLS	Adj. R-squared:	0.997
Method:	Least Squares	F-statistic:	2896.
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	1.70e-77
Time:	22:34:27	Log-Likelihood:	-35.442
No. Observations:	72	AIC:	86.88
Df Residuals:	64	BIC:	105.1
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	12.2698	0.140	87.667	0.000	11.990	12.549
C(prompt_type)[T.simple]	-0.0274	0.099	-0.277	0.783	-0.225	0.170
C(prompt_order)[T.prompt-first]	-0.0076	0.099	-0.077	0.939	-0.205	0.190
C(llm_tokens)[T.128]	0.0899	0.121	0.742	0.461	-0.152	0.332
C(llm_tokens)[T.256]	0.0621	0.121	0.512	0.610	-0.180	0.304
C(image_size)[T.256]	1.2720	0.121	10.494	0.000	1.030	1.514
C(image_size)[T.512]	8.4050	0.121	69.344	0.000	8.163	8.647
C(gene_type)[T.sequential]	11.9910	0.099	121.163	0.000	11.793	12.189

Omnibus:	22.341	Durbin-Watson:	2.245
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36.577
Skew:	1.162	Prob(JB):	1.14e-08
Kurtosis:	5.606	Cond. No.	5.89

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

ANOVA results for pipeline_time:

	sum_sq	df	F	PR(>F)
C(prompt_type)	1.206702e+04	1.0	1.051403	3.090442e-01
C(prompt_order)	5.917258e+04	1.0	5.155725	2.654382e-02
C(llm_tokens)	6.291738e+05	2.0	27.410055	2.519306e-09
C(image_size)	5.469189e+06	2.0	238.266056	2.225723e-30
C(gene_type)	1.789540e+07	1.0	1559.231606	1.178155e-46
Residual	7.345320e+05	64.0	NaN	NaN

OLS Regression Results

Dep. Variable:	pipeline_time	R-squared:	0.970
Model:	OLS	Adj. R-squared:	0.967
Method:	Least Squares	F-statistic:	299.5
Date:	Thu, 12 Dec 2024	Prob (F-statistic):	2.28e-46
Time:	22:34:27	Log-Likelihood:	-434.46
No. Observations:	72	AIC:	884.9
Df Residuals:	64	BIC:	903.1
Df Model:	7		
Covariance Type:	nonrobust		

```
==============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                       -79.0169     35.710     -2.213      0.030    -150.357      -7.677
C(prompt_type)[T.simple]        -25.8919     25.251     -1.025      0.309     -76.337      24.553
C(prompt_order)[T.prompt-first]  57.3356     25.251      2.271      0.027       6.891     107.780
C(llm_tokens)[T.128]            100.7735     30.926      3.259      0.002      38.992     162.555
C(llm_tokens)[T.256]            228.4510     30.926      7.387      0.000     166.669     290.233
C(image_size)[T.256]            349.5493     30.926     11.303      0.000     287.767     411.331
C(image_size)[T.512]            674.9604     30.926     21.825      0.000     613.178     736.742
C(gene_type)[T.sequential]      997.0902     25.251     39.487      0.000     946.645    1047.535
==============================================================================
Omnibus:                        7.021   Durbin-Watson:                   2.593
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                2.890
Skew:                          -0.155   Prob(JB):                        0.236
Kurtosis:                       2.068   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

------------------------------------------------------------------------------

ANOVA results for llm_metrics_vocab_diversity:
                   sum_sq    df         F    PR(>F)
C(prompt_type)   0.009322   1.0  4.116651  0.046627
C(prompt_order)  0.005742   1.0  2.535752  0.116221
C(llm_tokens)    0.001476   2.0  0.325881  0.723082
C(image_size)    0.004864   2.0  1.073993  0.347718
C(gene_type)     0.000028   1.0  0.012442  0.911533
Residual         0.144923  64.0       NaN       NaN
                            OLS Regression Results
==============================================================================
Dep. Variable:     llm_metrics_vocab_diversity   R-squared:                       0.129
Model:                                     OLS   Adj. R-squared:                  0.034
Method:                          Least Squares   F-statistic:                     1.352
Date:                         Thu, 12 Dec 2024   Prob (F-statistic):              0.241
Time:                                 22:34:27   Log-Likelihood:                121.33
No. Observations:                           72   AIC:                            -226.7
Df Residuals:                               64   BIC:                            -208.5
Df Model:                                    7
Covariance Type:                     nonrobust
==============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                        0.6033      0.016     38.036      0.000       0.572       0.635
C(prompt_type)[T.simple]        -0.0228      0.011     -2.029      0.047      -0.045      -0.000
C(prompt_order)[T.prompt-first]  0.0179      0.011      1.592      0.116      -0.005       0.040
C(llm_tokens)[T.128]            -0.0060      0.014     -0.438      0.663      -0.033       0.021
C(llm_tokens)[T.256]             0.0051      0.014      0.369      0.713      -0.022       0.033
C(image_size)[T.256]             0.0197      0.014      1.431      0.157      -0.008       0.047
C(image_size)[T.512]             0.0136      0.014      0.991      0.325      -0.014       0.041
C(gene_type)[T.sequential]       0.0013      0.011      0.112      0.912      -0.021       0.024
==============================================================================
Omnibus:                        1.032   Durbin-Watson:                   2.099
Prob(Omnibus):                  0.597   Jarque-Bera (JB):                0.513
Skew:                          -0.161   Prob(JB):                        0.774
Kurtosis:                       3.260   Cond. No.                         5.89
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--------------------------------------------------------------------------------

ANOVA results for llm_metrics_avg_sentence_length:
                     sum_sq    df          F    PR(>F)
C(prompt_type)     8.374657   1.0   0.823802  0.367476
C(prompt_order)  154.424040   1.0  15.190459  0.000235
C(llm_tokens)     28.643347   2.0   1.408801  0.251913
C(image_size)      0.613512   2.0   0.030175  0.970289
C(gene_type)      10.975480   1.0   1.079641  0.302687
Residual         650.614883  64.0        NaN       NaN
                                OLS Regression Results
==============================================================================================
Dep. Variable:     llm_metrics_avg_sentence_length   R-squared:                       0.238
Model:                                         OLS   Adj. R-squared:                  0.154
Method:                              Least Squares   F-statistic:                     2.853
Date:                             Thu, 12 Dec 2024   Prob (F-statistic):             0.0118
Time:                                     22:34:27   Log-Likelihood:                -181.41
No. Observations:                               72   AIC:                             378.8
Df Residuals:                                   64   BIC:                             397.0
Df Model:                                        7
Covariance Type:                         nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
Intercept                           17.3133      1.063     16.290      0.000      15.190      19.436
C(prompt_type)[T.simple]             0.6821      0.752      0.908      0.367      -0.819       2.183
C(prompt_order)[T.prompt-first]      2.9290      0.752      3.897      0.000       1.428       4.430
C(llm_tokens)[T.128]                 1.3426      0.920      1.459      0.150      -0.496       3.181
C(llm_tokens)[T.256]                 1.3333      0.920      1.449      0.152      -0.505       3.172
C(image_size)[T.256]                 0.2176      0.920      0.236      0.814      -1.621       2.056
C(image_size)[T.512]                 0.0556      0.920      0.060      0.952      -1.783       1.894
C(gene_type)[T.sequential]          -0.7809      0.752     -1.039      0.303      -2.282       0.720
==============================================================================
Omnibus:                        2.901   Durbin-Watson:                   2.564
Prob(Omnibus):                  0.234   Jarque-Bera (JB):                2.208
Skew:                           0.408   Prob(JB):                        0.332
Kurtosis:                       3.264   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--------------------------------------------------------------------------------

ANOVA results for llm_metrics_perplexity:
                     sum_sq    df          F        PR(>F)
C(prompt_type)     0.476034   1.0   0.041203  8.397912e-01
C(prompt_order)  342.843336   1.0  29.674489  8.698325e-07
C(llm_tokens)     21.890144   2.0   0.947341  3.931403e-01
C(image_size)     13.606832   2.0   0.588863  5.579361e-01
C(gene_type)      28.111834   1.0   2.433194  1.237223e-01
Residual         739.422117  64.0        NaN           NaN
                            OLS Regression Results
==============================================================================
Dep. Variable:     llm_metrics_perplexity   R-squared:                       0.355
Model:                                OLS   Adj. R-squared:                  0.284
Method:                     Least Squares   F-statistic:                     5.032
Date:                    Thu, 12 Dec 2024   Prob (F-statistic):           0.000138
Time:                            22:34:27   Log-Likelihood:                -186.01
```

```
No. Observations:                    72   AIC:                             388.0
Df Residuals:                        64   BIC:                             406.2
Df Model:                             7
Covariance Type:                nonrobust
==========================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                       26.5895      1.133     23.468      0.000      24.326      28.853
C(prompt_type)[T.simple]        -0.1626      0.801     -0.203      0.840      -1.763       1.438
C(prompt_order)[T.prompt-first]  4.3643      0.801      5.447      0.000       2.764       5.965
C(llm_tokens)[T.128]             0.9549      0.981      0.973      0.334      -1.005       2.915
C(llm_tokens)[T.256]             1.3046      0.981      1.330      0.188      -0.656       3.265
C(image_size)[T.256]             1.0480      0.981      1.068      0.290      -0.912       3.008
C(image_size)[T.512]             0.6875      0.981      0.701      0.486      -1.273       2.648
C(gene_type)[T.sequential]      -1.2497      0.801     -1.560      0.124      -2.850       0.351
==============================================================================
Omnibus:                        1.959   Durbin-Watson:                   2.076
Prob(Omnibus):                  0.375   Jarque-Bera (JB):                1.389
Skew:                           0.325   Prob(JB):                        0.499
Kurtosis:                       3.199   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--------------------------------------------------------------------------------

ANOVA results for story_context_fid_context_fid:
                   sum_sq    df         F    PR(>F)
C(prompt_type)   0.005105   1.0  2.231423  0.140143
C(prompt_order)  0.007847   1.0  3.430022  0.068635
C(llm_tokens)    0.001831   2.0  0.400074  0.671935
C(image_size)    0.000161   2.0  0.035100  0.965527
C(gene_type)     0.001914   1.0  0.836500  0.363833
Residual         0.146422  64.0       NaN       NaN
                            OLS Regression Results
==============================================================================================
Dep. Variable:     story_context_fid_context_fid   R-squared:                       0.103
Model:                                       OLS   Adj. R-squared:                  0.005
Method:                            Least Squares   F-statistic:                     1.053
Date:                           Thu, 12 Dec 2024   Prob (F-statistic):              0.404
Time:                                   22:34:27   Log-Likelihood:                120.96
No. Observations:                             72   AIC:                            -225.9
Df Residuals:                                 64   BIC:                            -207.7
Df Model:                                      7
Covariance Type:                       nonrobust
==========================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                        0.0946      0.016      5.932      0.000       0.063       0.126
C(prompt_type)[T.simple]        -0.0168      0.011     -1.494      0.140      -0.039       0.006
C(prompt_order)[T.prompt-first] -0.0209      0.011     -1.852      0.069      -0.043       0.002
C(llm_tokens)[T.128]            -0.0114      0.014     -0.824      0.413      -0.039       0.016
C(llm_tokens)[T.256]            -0.0015      0.014     -0.110      0.913      -0.029       0.026
C(image_size)[T.256]            -0.0019      0.014     -0.138      0.891      -0.029       0.026
C(image_size)[T.512]             0.0018      0.014      0.127      0.899      -0.026       0.029
C(gene_type)[T.sequential]       0.0103      0.011      0.915      0.364      -0.012       0.033
==============================================================================
Omnibus:                       15.831   Durbin-Watson:                   2.082
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               18.293
Skew:                           1.061   Prob(JB):                     0.000107
```

```
Kurtosis:                       4.262   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

------------------------------------------------------------------------------

ANOVA results for story_context_fid_avg_similarity:
                   sum_sq    df          F    PR(>F)
C(prompt_type)   0.000032   1.0   0.002505  0.960238
C(prompt_order)  0.202557   1.0  15.904994  0.000174
C(llm_tokens)    0.023219   2.0   0.911609  0.407031
C(image_size)    0.032637   2.0   1.281339  0.284691
C(gene_type)     0.000260   1.0   0.020434  0.886781
Residual         0.815067  64.0        NaN       NaN
                            OLS Regression Results
==============================================================================================
Dep. Variable:     story_context_fid_avg_similarity   R-squared:                       0.241
Model:                                          OLS   Adj. R-squared:                  0.158
Method:                               Least Squares   F-statistic:                     2.902
Date:                              Thu, 12 Dec 2024   Prob (F-statistic):            0.0107
Time:                                      22:34:27   Log-Likelihood:                 59.158
No. Observations:                                72   AIC:                            -102.3
Df Residuals:                                    64   BIC:                            -84.10
Df Model:                                         7
Covariance Type:                          nonrobust
==============================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                        0.4732      0.038     12.578      0.000       0.398       0.548
C(prompt_type)[T.simple]        -0.0013      0.027     -0.050      0.960      -0.054       0.052
C(prompt_order)[T.prompt-first]  0.1061      0.027      3.988      0.000       0.053       0.159
C(llm_tokens)[T.128]            -0.0146      0.033     -0.448      0.655      -0.080       0.050
C(llm_tokens)[T.256]             0.0286      0.033      0.879      0.383      -0.036       0.094
C(image_size)[T.256]             0.0431      0.033      1.322      0.191      -0.022       0.108
C(image_size)[T.512]            -0.0039      0.033     -0.120      0.904      -0.069       0.061
C(gene_type)[T.sequential]       0.0038      0.027      0.143      0.887      -0.049       0.057
==============================================================================
Omnibus:                        5.817   Durbin-Watson:                   2.212
Prob(Omnibus):                  0.055   Jarque-Bera (JB):                5.038
Skew:                           0.613   Prob(JB):                       0.0805
Kurtosis:                       3.419   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

------------------------------------------------------------------------------

ANOVA results for image_metrics_avg_ssim:
                   sum_sq    df          F        PR(>F)
C(prompt_type)   0.000196   1.0   0.226629  6.356562e-01
C(prompt_order)  0.000444   1.0   0.514673  4.757334e-01
C(llm_tokens)    0.000526   2.0   0.304720  7.383935e-01
C(image_size)    0.082900   2.0  48.040350  1.815148e-13
C(gene_type)     0.001329   1.0   1.540323  2.190997e-01
Residual         0.055220  64.0        NaN           NaN
                            OLS Regression Results
==============================================================================
Dep. Variable:     image_metrics_avg_ssim   R-squared:                       0.607
```

```
Model:                                OLS   Adj. R-squared:                  0.564
Method:                     Least Squares   F-statistic:                     14.14
Date:                    Thu, 12 Dec 2024   Prob (F-statistic):           6.15e-11
Time:                            22:34:27   Log-Likelihood:                 156.07
No. Observations:                      72   AIC:                            -296.1
Df Residuals:                          64   BIC:                            -277.9
Df Model:                               7
Covariance Type:                nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                        0.1030      0.010     10.524      0.000       0.083       0.123
C(prompt_type)[T.simple]         0.0033      0.007      0.476      0.636      -0.011       0.017
C(prompt_order)[T.prompt-first]  0.0050      0.007      0.717      0.476      -0.009       0.019
C(llm_tokens)[T.128]            -0.0036      0.008     -0.422      0.675      -0.021       0.013
C(llm_tokens)[T.256]             0.0030      0.008      0.358      0.722      -0.014       0.020
C(image_size)[T.256]             0.0286      0.008      3.375      0.001       0.012       0.046
C(image_size)[T.512]             0.0819      0.008      9.657      0.000       0.065       0.099
C(gene_type)[T.sequential]      -0.0086      0.007     -1.241      0.219      -0.022       0.005
==============================================================================
Omnibus:                        9.350   Durbin-Watson:                   1.889
Prob(Omnibus):                  0.009   Jarque-Bera (JB):                9.088
Skew:                           0.833   Prob(JB):                       0.0106
Kurtosis:                       3.501   Cond. No.                         5.89
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

--------------------------------------------------------------------------------

ANOVA results for image_metrics_diversity:
                       sum_sq    df          F        PR(>F)
C(prompt_type)   5.595441e+05   1.0   0.059033  8.088075e-01
C(prompt_order)  2.936267e+07   1.0   3.097835  8.317470e-02
C(llm_tokens)    6.645229e+05   2.0   0.035054  9.655714e-01
C(image_size)    6.718554e+08   2.0  35.441212  4.357000e-11
C(gene_type)     1.447588e+07   1.0   1.527241  2.210438e-01
Residual         6.066207e+08  64.0        NaN           NaN
                            OLS Regression Results
==============================================================================
Dep. Variable:     image_metrics_diversity   R-squared:                       0.542
Model:                                OLS   Adj. R-squared:                  0.492
Method:                     Least Squares   F-statistic:                     10.81
Date:                    Thu, 12 Dec 2024   Prob (F-statistic):           6.59e-09
Time:                            22:34:27   Log-Likelihood:                -676.25
No. Observations:                      72   AIC:                             1368.
Df Residuals:                          64   BIC:                             1387.
Df Model:                               7
Covariance Type:                nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      3612.6062   1026.236      3.520      0.001    1562.464    5662.749
C(prompt_type)[T.simple]        176.3116    725.659      0.243      0.809   -1273.358    1625.981
C(prompt_order)[T.prompt-first] -1277.2077    725.659     -1.760      0.083   -2726.877     172.462
C(llm_tokens)[T.128]            -233.0762    888.747     -0.262      0.794   -2008.552    1542.399
C(llm_tokens)[T.256]             -88.4435    888.747     -0.100      0.921   -1863.919    1687.032
C(image_size)[T.256]            1916.3768    888.747      2.156      0.035     140.901    3691.852
C(image_size)[T.512]            7222.0998    888.747      8.126      0.000    5446.624    8997.575
C(gene_type)[T.sequential]       896.7806    725.659      1.236      0.221    -552.889    2346.450
```

Omnibus:	18.482	Durbin-Watson:	1.945
Prob(Omnibus):	0.000	Jarque-Bera (JB):	31.556
Skew:	0.924	Prob(JB):	1.41e-07
Kurtosis:	5.666	Cond. No.	5.89

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.