ViStoryBench: Comprehensive Benchmark Suite for Story Visualization – Supplementary Materials –

Appendix

A. Broader Limitations and Societal Impact

The design and scope of ViStoryBench are subject to several limitations, which reflect current technological constraints and evaluation challenges in the field of story-oriented generation:

Focus on multi-image consistency: While the long-term objective of this research community is synchronized audio-visual storytelling with full scene dynamics, the current benchmark is deliberately scoped to multi-image generation with an emphasis on inter-shot consistency. This focus allows us to address the most immediate challenges in visual narrative coherence without introducing additional complexity from temporal modeling or audio alignment, which are active areas of research in their own right. Since established benchmarks like VBench [41] already provide comprehensive metrics for single-shot video temporal modeling, we avoid redundant efforts in this area.

Lack of background-reference evaluation: Due to limited support for background-conditioned generation in current open-source story visualization models, ViStoryBench does not incorporate background reference images or include scene-level image-to-image similarity evaluation, with only a minimal number of single-scene stories included. Instead, scene descriptions are provided via text prompts, and scene consistency is assessed through prompt alignment. Future work will expand the dataset to include more single-scene multi-shot stories along with corresponding scene reference images to enhance the scope and functionality of the benchmark.

Inherent trade-offs in evaluation metrics: The benchmark employs a hybrid evaluation strategy that combines expert models (for stability and continuity) and VLM-based scorers (for semantic richness and narrative alignment). However, this approach involves fundamental compromises: expert models may lack adaptability in visually or narratively complex scenarios, while VLMs remain susceptible to hallucination and may not always align with human perceptual judgments. These trade-offs underline the lack of a universally optimal automated metric for all aspects of story visualization.

Dataset Limitations and Copyright Concerns. Some images in our dataset are derived from well-known movies, TV shows, and animations. While these samples are used solely for academic research and benchmarking purposes, they may raise copyright concerns. We do not claim ownership of any copyrighted material, and all third-party content is included under fair use principles for non-commercial research and analysis. Nevertheless, users of our dataset should be aware of potential legal constraints when repurposing or redistributing the data. Additionally, the inclusion of well-known visual content may cause certain metrics to become overfitted to these familiar styles or characters, potentially leading to metric manipulation or over-optimization.

Language Sensitivity. Our benchmark supports both Chinese and English story prompts. While we select the appropriate language for each model based on its design and documentation, we do not control for potential discrepancies in generation quality caused by language differences. This may introduce variation that is not attributable to model capability alone.

Scope of Evaluation. Our benchmark currently does not support accurate evaluation of comic-style or manga generation tasks that involve multi-panel layouts within a single image, due to the lack of a robust panel segmentation method. Similarly, we do not assess inference efficiency or runtime performance across models. For video-based story generation methods, our benchmark does not yet include a comprehensive evaluation of temporal coherence, frame-level consistency, or other video-specific quality metrics. These remain important future directions.

Societal Impact. We envision story visualization models as promising tools for education, creativity, and cultural

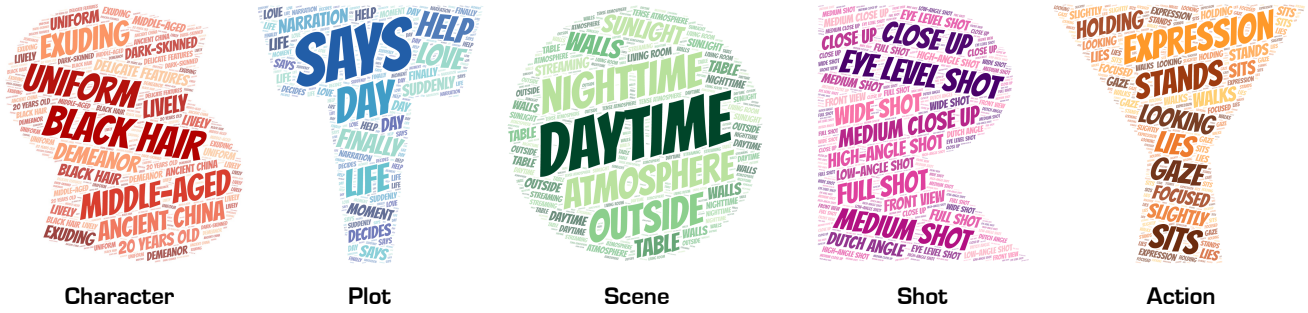


Figure S1. **Words Cloud**. Visualization of narrative elements from the stories in our ViStoryBench dataset: spanning character traits (e.g., black hair, middle-aged, uniform), plot points (e.g., says, finally, love), scene settings (e.g., daytime, atmosphere, outside, sunlight, living room), shot types (e.g., eye-level, close-up, high-angle), and character actions (e.g., expression, stands, gaze).

preservation. In curating the dataset, we made conscious efforts to include diverse narratives from multiple cultures and regions. However, generative models are still susceptible to reproducing stereotypes and amplifying data-driven biases. It is vital that these tools are developed and deployed responsibly.

Conclusion. Despite extensive efforts to align metric selection with the requirements of each evaluation dimension, these limitations reflect persistent challenges in the current evaluation paradigm. They also indicate meaningful directions for future work in developing more comprehensive and reliable story visualization benchmarks.

Finally, we emphasize that generative models should not be used to create or disseminate false or misleading content. Addressing such risks requires active collaboration between researchers, platform providers, and policymakers to ensure safe and ethical applications.

B. Overall Insights

Based on the table of automated test results on ViStoryBench and ViStoryBench-Lite, we have analyzed the performance of various story visualization methods across multiple metrics. Here are the key insights and patterns observed from a research perspective:

Performance of Multi-modal Large Models is Dominant, Especially GPT-4o.

① GPT-4o [15] achieves the best or second-best performance in critical metrics like Alignment Score (3.673), OCCM Score (93.5), CIDS (Cross: 0.571 and Self: 0.679), and CSD (Cross: 0.481 and Self: 0.680), indicating strong narrative understanding, character consistency and style similarity. This suggests that large-scale multi-modal models excel at high-level semantic alignment and coherence, likely due to their extensive pre-training on diverse data.

② However, GPT-4o [15] lags in Inception Score (9.02) and Aesthetics Score (5.49), which measure image diversity and visual quality. This implies a trade-off: while LLMs handle narrative complexity well, they may struggle with low-level visual fidelity compared to specialized methods.

Commercial Software Shows Strengths in Visual Quality but Inconsistencies in Narrative Tasks.

① Commercial tools like MorphicStudio [23] excel in style consistency (CSD_Cross: 0.653), while Doubao [1] perform well in alignment (3.494). This highlights their optimization for production-ready output.

② However, they exhibit variability: for example, MOKI [22] has high aesthetics (5.79) but poor character consistency (CIDS_Cross: 0.214). This indicates that commercial tools may prioritize aesthetic appeal over fine-grained narrative controls, leading to imbalances in evaluation dimensions.

Story Image Methods Excel in Specific Niches, but Lack Uniformity.

① Methods like OmniGen2 [35] lead in character consistency (CIDS_Self: 0.537) and OCCM (90.8), demonstrating strengths in maintaining identity across frames. Story-Adapter [20] variants achieve high style consistency (CSD_Cross: 0.456), showing progress in specialized tasks.

② However, performance varies widely: SEED-Story [39] and TheaterGen [2] score low on multiple metrics (e.g., Alignment <2.00 and CIDS_Cross <0.35), indicating that some methods overfit to specific scenarios or lack generalization. The reliance on reference images (e.g., Story-Adapter [20] with image-ref) often boosts consistency but may limit creativity.

Story Video Methods Underperform in Key Metrics, Revealing Challenges in Temporal Modeling.

① Video-based Methods like Vlogger [45] and MovieAgent [34] generally score lower in style and character consistency (e.g., CSD_Cross <0.3) compared to image-based methods. This suggests that temporal modeling introduces additional complexity, hindering per-frame quality.

② An exception is MovieAgent [34] (SD3), which achieves strong alignment (3.16), implying that leveraging advanced image-based diffusion models (e.g., SD3 [8]) can mitigate some issues. Yet, overall, video methods lag in metrics like Inception Score, indicating limited diversity in generated sequences.

Excellent Performance in Multi-shot Video Models, Especially Sora2.

① Multi-shot video generation models excel in cross-shot character and scene consistency, as evidenced by high self-similarity score (e.g., CIDS_Self 0.813 and CSD_Self 0.713 for Sora2 [24]), outperforming many image-based methos. This likely results from training on large-scale character-consistent multi-shot movie data. In contrast, most image-based methods lack specialized training on multi-shot storyboard data.

② However, current multi-shot video models struggle with visual reference adherence, indicated by relatively lower cross-similarity scores (e.g., CIDS_Cross 0.738 and CSD_Cross 0.515 for Sora2 [24]), revealing a gap in style and scene constraint preservation compared to reference-based image methods.

Trade-offs Between Consistency and Creativity Are Evident. Methods with high character consistency and high copy-paste score (lower is better) often have lower diversity (e.g., GPT-4o). This underscores a fundamental tension in story visualization: optimizing for one dimension may compromise another.

In story visualization tasks, comprehensive evaluation metrics are extremely important. For instance, the simple Copy-Paste Baseline achieves optimal results across numerous metrics. However, its alignment score is notably low. Although IS can generally measure the quality and diversity of image generation, it is quite challenging to compare different models by examining the IS metric alone. When using only text as input, StoryDiffusion [42] and Story-Adapter [20] achieve excellent IS and aesthetic quality. However, relying solely on text input clearly cannot produce results that resemble the features and styles of the character reference images.

Our quantitative metrics demonstrate alignment with qualitative observations. For Story-Adapter [20], the scoring consistency between automated metrics and human evaluation is particularly evident: (1) In text-only mode (its native setting), the overall quality score (scale=5) systematically surpasses the baseline (scale=0), as theoretically expected; (2) When using image references, scale=0 achieves higher cross-similarity but lower self-similarity compared to scale=5 in both CIDS and CSD.

ViStoryBench-Lite Reveals Real-World Gaps Results on ViStoryBench-Lite (focused on practical scenarios) show that commercial and LLM methods perform well in alignment but struggle with low-level metrics (e.g., Gemini-2.0 [6] has Align: 3.150 but CSD_Cross: 0.361). This indicates that real-world applications require balancing semantic and visual qualities, and current methods may not fully address this.

Conclusion and Future Directions:

① No single method dominates all metrics, emphasizing the need for task-specific model selection. For narrative-heavy tasks, LLMs like GPT-4o [15] are preferable; for visual quality, commercial tools or specialized image methods may suffice.

② Future work should focus on hybrid approaches: integrating LLMs for planning with diffusion models for visual quality, and improving evaluation metrics to better capture storytelling aspects like pacing and emotion.

③ A promising direction for story visualization is to combine the temporal coherence of video models with the precise reference alignment of image-based generators, enabling customized multi-shot output with both consistency and control.

④ The variability in results underscores the value of benchmarks like ViStoryBench for guiding progress. Researchers should prioritize methods that balance consistency, diversity, and alignment.

C. Details of Dataset Collection and Statistics

Story visualization datasets have shown notable growth and evolution in terms of scale, image resolution, automated generation pipelines, and stylistic diversity—reflecting both technological advancements and the expanding scope of research interests. We summarize story visualization datasets [13, 14, 16, 18, 39, 40] in Table S3. A key distinction of ViStoryBench compared to existing datasets lies in our construction methodology: rather than extracting captions from visual keyframes to construct a narrative, we take a top-down approach by generating structured shot descriptions directly from full textual stories. Additionally, when curating benchmarks, we place particular emphasis on en-

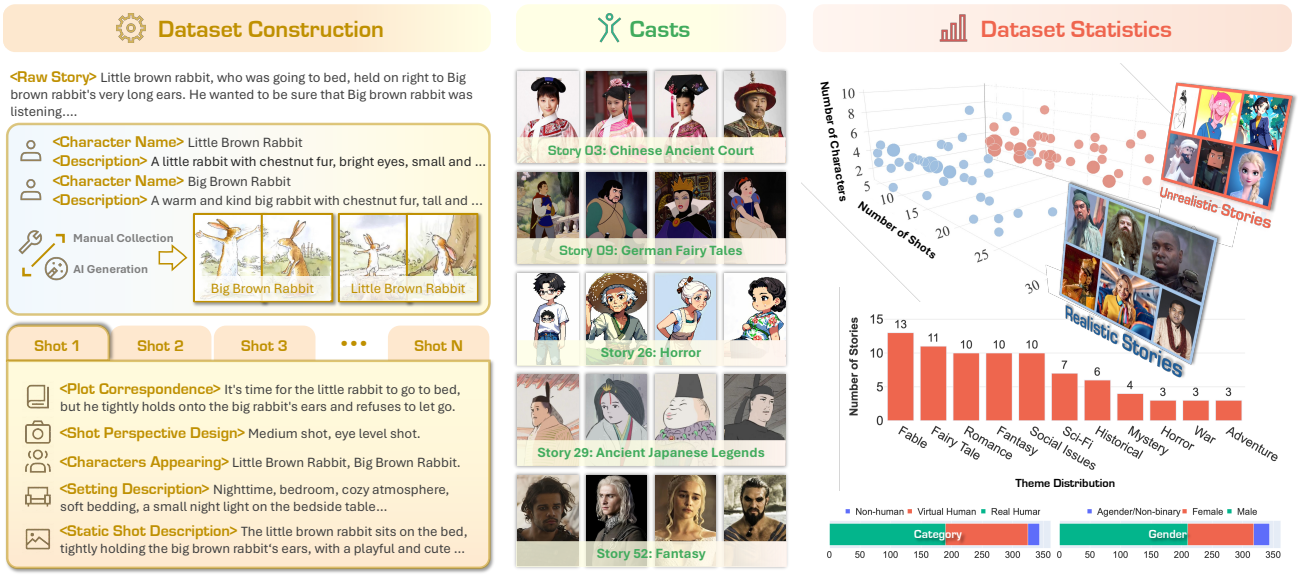


Figure S2. **Overview of ViStoryBench Dataset.** Dataset Construction: We build a story generation pipeline powered by large language models (LLMs), followed by human verification to ensure quality and consistency. Casts: Reference images for characters are manually curated to maintain a consistent visual style. Dataset Statistics: ViStoryBench dataset exhibits a broad distribution across story categories, stylistic variations, and character diversity, enabling comprehensive evaluation of storytelling generation models.

suring a broad range of styles and thematic diversity. Table S1 presents a detailed breakdown of the 10 visual styles curated in our dataset, which are classified based on their character reference images.

Table S1. Distribution of the 10 Visual Styles in ViStoryBench. The classification is based on the visual style of the character reference images for each story.

Style Category	Number of Stories
Photorealistic / Live-action photo	39
Anime / Cel-shading style	14
Children's book / Cartoon	7
Classic fairy tale illustration / Vintage	5
Classical oil painting / Religious art	4
Flat vector illustration	4
3D / Voxel / Claymation style	3
Chinese ink painting / Line art / Silhouette	2
Action manga / American comics style	1
CG realistic fantasy illustration	1
Total	**80**

The collected dataset spans a wide array of genres, including 13 folktales, 10 romance stories, 4 suspense/crime stories, 3 horror narratives, 6 historical tales, 10 fantasy stories, 7 science fiction stories, 3 war stories, 10 stories about social life, 3 survival/adventure stories, and 11 fairy tales. These stories are segmented and adapted into detailed shot scripts using the in-house LLM model. The full dataset contains 1,317 shots in total, with each story comprising between 4 and 30 shots (averaging 16.5 shots per story). Basic statistics shown in Figure S2

To support a wide range of methods, all test-related textual prompts are provided in both English and Chinese. For methods that only support Chinese, or perform significantly better with Chinese input, we use the Chinese version; otherwise, English inputs are used. Each individual shot is annotated with five structured fields: *Setting Description*, *Plot Correspondence*, *Onstage Characters*, *Static Shot Description*, and *Shot Perspective Design*.

In our curation process, we made a conscious effort to incorporate narratives from diverse cultural backgrounds. The cultural origins of the 80 stories are distributed across several major regions, including Chinese (39 stories), Euro-American (27), Japanese (8), African (3), Islamic/Middle Eastern (2), and Indian (1). This diversity is complemented by a wide array of thematic genres: the dataset includes 13 folktales, 10 romance stories, 4 suspense/crime stories, 3 horror narratives, 6 historical tales, 10 fantasy stories, 7 science fiction stories, 3 war stories, 10 stories about social life, 3 survival/adventure stories, and 11 fairy tales.

C.1. Character Reference Image

For most well-known stories in our dataset, character reference images are directly sourced from existing visual works such as movies, animated films, or television series. For lesser-known or original stories, we adopt one of two strategies to obtain reference images for the main characters: (1) retrieving representative screenshots from films or TV shows with similar settings and styles (covering 16 stories),

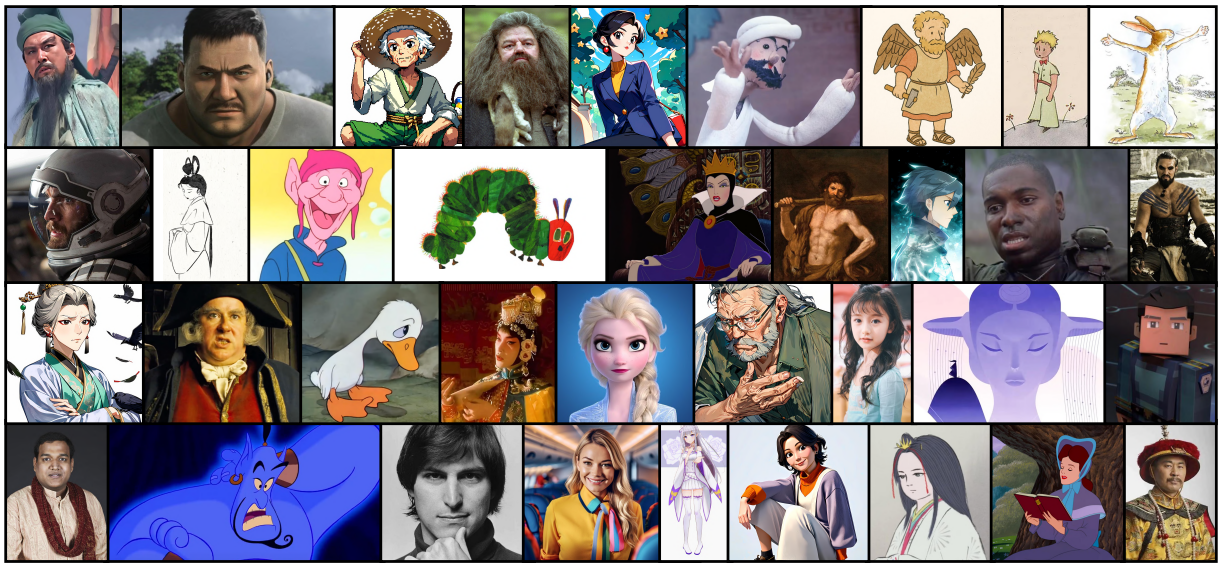


Figure S3. **Random Character Reference Samples from the Dataset**. The reference images include full-body shots, half-body shots, or portraits, spanning diverse visual styles from photorealistic to a variety of animation-inspired designs.

or (2) using the SDXL model [27] to generate high-quality stylized animation portraits (covering 7 stories).

In total, our dataset includes **344 unique characters**, which can be categorized into **190 real humans**, **135 virtual humans** (e.g., animated or game characters), and **19 non-human entities** (e.g., animals or creatures). Regarding gender annotation, there are **210 male**, **108 female**, and **26 characters** who are either genderless or non-binary. Each character is associated with between 1 and 10 reference images, resulting in a total of **509 reference images**, with 89 characters having more than one image. A selection of these reference images is visualized in Figure S3.

Furthermore, we categorize all 80 stories into two distinct types based on the visual style of the main character references: **realistic stories** and **unrealistic stories**. The realistic category includes 39 stories whose characters are portrayed using photographic or cinematic images. These characters are additionally labeled with ethnicity information following prior works [3, 4, 26]. The remaining 41 stories are labeled as unrealistic, typically involving animation, stylized art, or fantasy characters.

This classification allows us to conduct stratified evaluations and analyze how different generation methods perform across story types with distinct visual and semantic characteristics.

C.2. Prompts for Dataset Construction

To automate the transformation of story narratives into detailed multi-shot descriptions, we leverage a large language model (LLM) as a shot planner. This LLM is tasked with segmenting each narrative into a coherent sequence of visual shots, ensuring consistency in character presence, camera composition, environment description, and story progression.

We propose a structured prompt engineering approach for data generation, which systematically decomposes the complex story visualization task into well-defined, verifiable subtasks to achieve precise control over LLM outputs. This methodology aligns with rigorous benchmark construction standards in academic research and offers a valuable technical framework for building complex, structured multimodal datasets. Our approach converts an MLLM into a controllable visual narrative script generator and employs five core strategies:

① **Multi-grained task decomposition**, breaking the task into five structured modules—Plot Correspondence, Setting Description, Shot Perspective Design, Onstage Characters, and Static Shot Description—enabling the LLM to focus on simpler subproblems for improved accuracy and stability.

② **Professional knowledge infusion**, incorporating cinematic expertise such as standardized shot terminology (e.g., Wide Shot, Low-angle) and narrative principles (e.g., Smooth Narrative Transitions, Emotional Resonance).

③ **Multi-dimensional information isolation**, enforcing modality separation (e.g., excluding characters from setting descriptions) to prevent spurious correlations and support combinatory generalization.

④ **Visual-friendly description**, ensuring all textual content is concrete and directly depictable (e.g., avoiding abstract expressions in favor of visually grounded de-

Table S2. **Comparison between Lite and Full Settings across Multiple Evaluation Dimensions**. Quantitative comparison between the full dataset and the lite subset under a single method. The close alignment across all metrics, including style and character consistency, generative quality, diversity, and prompt alignment, which demonstrates that the lite subset serves as a representative proxy for the full dataset, enabling efficient yet reliable evaluation.

	Style Consistency		**Character Consistency**		**Character Matching**	**Generative Quality**	**Diversity**	**Prompt Alignment**			
	Cross	Self	Cross	Self	OCCM	Aesthetic	Inception	Scene	Shot	CI	IA
Full	0.325	0.569	0.427	0.600	62.487	4.894	11.510	2.331	2.891	2.176	2.101
Lite	0.373	0.626	0.469	0.620	65.351	5.027	9.787	2.779	2.965	2.564	2.290
Diff	13.78%	9.55%	9.34%	3.21%	4.48%	2.68%	16.17%	17.57%	2.52%	16.34%	8.62%

Table S3. **Comparison of Multi-modal Storytelling Datasets**. Representative story datasets are summarized in terms of annotation method, image scale, resolution, number of shots per story, and visual style.

Datasets	**Caption**	**# Images**	**Resolution**	**# Shots**	**Style**
VIST [14]	Manual	146k	-	5	Realistic
Flintstones [13]	Manual	123k	128×128	5	Anime
Pororo [16]	Manual	74k	128×128	5	Anime
StorySalon [18]	ASR	160k	432×803	14	Anime
StoryStream [39]	Generated	258k	480×854	30	Anime
OpenStory [40]	Generated	107M	720p & 1080p	28	Realistic

scriptions).

⑤ **Contextual narrative modeling**, maintaining coherence across shot sequences by considering preceding and subsequent shots. Thus, our prompt set defines an LLM-powered automated data generation framework, enabling efficient construction of a high-quality, consistent, and domain-informed benchmark for visual storytelling.

Below, we present the specific system prompt utilized for this purpose.

D. Qualitative Results

We provide the visualization generation results of the methods tested on **Story 09** and **Story 01**, as shown in Figure S4 and Figure S5. We sample the first five shots of the story for display to offer a concise yet representative comparison. At the top of the figure, we include a "Copy-Paste Baseline" that visually presents the ground-truth character presence in each frame using manually cropped reference images. This serves as a reference for evaluating the accuracy of character depiction across methods. Below the baseline, we showcase the generation results from all 18 evaluated open-source methods (including their key variants), as well as several leading commercial tools. These comparisons highlight differences in prompt alignment, character consistency, and visual quality across models.

To facilitate better examination of the visual outputs, we also provide a full frame-by-frame visualization of the entire story. (Due to policy restrictions, the demo video was recorded anonymously and submitted as supplementary material.)

This detailed visualization enables a more comprehensive evaluation of each method's performance over the complete narrative sequence, especially in terms of temporal coherence, character persistence, and scene transitions. We encourage readers to explore this link for in-depth visual analysis beyond the summarized frames shown in the main figure.

More results can be found on the website: https://huggingface.co/datasets/ViStoryBench/ViStoryBenchResult

E. ViStoryBench-Lite Benchmark

E.1. Effectiveness of ViStoryBench-Lite

To assess the representativeness and reliability of ViStoryBench-Lite, we conduct a comprehensive comparison between the full dataset and the **Lite** subset. We first perform a distributional analysis over story categories to examine the content diversity across both subsets. As shown in Figure **??**, the **Lite** subset exhibits a highly similar category distribution to the full dataset, suggesting a well-preserved narrative and visual diversity.

Further, we apply a unified generation method across both subsets and report quantitative results over all major evaluation dimensions. The detailed comparison is presented in Table S2, which includes metrics on style consistency, character consistency, generative quality, diversity, and prompt alignment. The results demonstrate that the performance difference between the **Lite** and full datasets is minimal across most dimensions. Notably, only marginal discrepancies are observed in certain VLM-based

Prompt for Dataset Construction (Part 1)

You are a seasoned film script artist skilled at transforming descriptive text from novel scripts into visual content descriptions. You also adapt scripts into static shot scripts. Your designs must incorporate a wide variety of compositions to perfectly capture the script's content through imagery, ensuring the storyline is effectively conveyed in the visual descriptions. In addition, you also need to provide a comprehensive introduction to the characters that appear in the shot scripts, mainly describing their appearance and clothing.

Task Description

You are required to write shot scripts based on the user's input story, totaling <num_of_shots> shots. Each shot can feature 0 to 3 characters, meaning it can be a scene shot, single-character shot, two-character shot, or three-character shot. For each shot in the shot script, you need to output <Plot Correspondence>, <Setting Description>, <Shot Perspective Design>, <Onstage characters>, <Static Shot Description>.
The <Plot Correspondence>section requires dividing the original plot into <num_of_shots> scenes, presenting the plot of each scene in the form of narration and dialogue. Note that when dividing the input plot into different scenes, the rationality of the plot needs to be considered.

Concept Explanation of Various Fields in Shot Scripts

- Setting Description: The story will be divided into <num_of_shots> scenes. The setting refers to the environmental setup of each scene. It should not include any characters. You need to describe all elements in the environment in detail in this field, so that the scene where the story takes place can be vividly recreated. Standard writing format: time, location, atmosphere description, other elements in the environment, lighting effects.
- Shot Perspective Design: Shot Perspective Design refers to information from several dimensions: shot distance, camera angle, and camera type.
- Onstage characters: Please select the characters appearing in this scene from the character list. The number of characters should be controlled between 0-3. If no characters appear, leave it blank.
- Static Shot Description: This part describes the static actions or positions of characters and items in the scene, ensuring that it describes a fixed state. Writing format: <character position>, <character expression>, <character action>, <position of elements in the scene>.

Requirements for Creating Setting Description in Shot Scripts

- You need to design a shot script for the user's input story. Break the story into <num_of_shots> main scenes and write scene descriptions for each of these <num_of_shots> scenes.
- Note that character descriptions should not appear in Setting Descriptions. Only describe the scene itself, ensuring that the scene is consistent with the original story. Do not include backgrounds, items, or other elements that are not present in the story.
- Think from the perspective of the visuals, using the visuals to drive the content of the shots. Ensure that all plot elements can be directly depicted through the visuals. Avoid thinking from the perspective of a screenwriter's script and refrain from using abstract or metaphorical expressions.
- Pay attention to the consistency between the characters' locations and the Setting Descriptions.
- When writing the background content, do not directly use the expressions from the origin story. Use clear and concise sentences to describe in detail all the elements included in the background visuals and their relationships.
- If there are characters in the visuals, clearly express their facial expressions, demeanor, and actions.
- Pay attention to the visual narrative continuity between adjacent shot panels.

Requirements for Creating Shot Perspective Design in Shot Scripts

- Continuity in Shot Composition: Adjacent shots should maintain coherence in shot composition and camera angles. By employing a diverse yet consistent combination of shot compositions and camera angles, create a viewing experience that is both spatially immersive and visually engaging.
- Shot Distance Selection: Choose the most appropriate shot distance from "wide shot, full shot, medium long shot, medium shot, medium close up, close up" based on the emotional atmosphere of the current scene. A smaller subject size results in a more relaxed emotional tone, while a larger subject size creates a tenser atmosphere. Consider the shot distance design of preceding and following shots as well.

Prompt for Dataset Construction (Part 2)

- Guidelines for Shot Composition Combinations:
  - Smooth Narrative Transitions: The way shot compositions are connected impacts narrative fluidity. Effective transitions involve gradual tightening or loosening. Moving from wider to tighter shots is called "tightening," while moving from tighter to wider shots is called "loosening." Avoid abrupt shifts from Wide Shot to extreme close-ups or vice versa.
  - Avoid Repetitive Compositions: Ensure that adjacent shots have different compositions to prevent visual monotony.
  - Emotional Resonance: Since shot composition affects emotional tone, match shot compositions with the emotional intensity of the plot. For instance, intense emotional scenes are better suited for tighter shots.
  - Rhythmic Pacing: The combination of shot compositions influences the visual rhythm of the scene. Use an appropriate mix of shot compositions within each episode to convey the narrative's pace effectively.
- Camera Angle Selection: Opt for the camera angle—"front view, side view, back view"—that best suits the current scene's visual requirements.
- Camera Type Selection: Choose the camera type—"Eye level shot, low-angle shot, high-angle shot, bird's eye view shot, dutch angle shot, foreshortening, inverted shot"—that aligns with the scene's content and emotional tone. Consider the camera types used in preceding and following shots for continuity.
- Camera Type Combinations: Select camera types that complement the scene's content and emotional context. Pay attention to how different camera types interact to enhance the visual storytelling.
- Please refer to the provided materials to choose the most fitting camera type for the current scene's content and emotional tone, ensuring that the combination of camera types supports the overall narrative effectively.

Reference Materials for Camera Design

For Shot Distance selection:

- Wide Shot: Displays the relationship between characters and their environment, commonly used to showcase scenes and background settings.
- Full shot: Shows the entire body of a character, often used to present full actions or the overall view of a scene.
- Medium long shot: Captures from above the character's knees.
- Medium shot: Captures from above the character's waist.
- Medium close up: Captures from above the character's chest.
- Close up: focus on a close-up of the character's head or face, with the background and environment typically blurred or entirely out of view.

For determining the relationship between Camera Type and content:

- Eye level shot: The camera is positioned at the same level as the eyeline.
- Low-angle shot: The camera shoots upward from below the eyeline, enhancing the subject's authority or size, often conveying power or intimidation. Suitable for emphasizing an individual's dominance or creating visual pressure, such as highlighting a hero or villain.
- High-angle shot: The camera looks down from above the eyeline, showing the breadth of a scene or diminishing the visual importance of the subject. Effectively reduces the visual scale of characters or objects, used to depict a sense of isolation or helplessness in characters, or to present vast landscapes.
- Bird's eye view shot: The camera shoots downward from a high altitude, providing a top-down perspective, usually covering extensive geographical areas. Highly effective when a global view of events or environments is needed.
- Dutch angle shot: The camera is deliberately tilted during filming, often used to create a sense of imbalance or tension. Particularly effective in portraying scenes of chaos, tension, or psychological instability.
- Foreshortening: Emphasizes depth of field through perspective techniques, making the relationship between foreground and background more prominent. Suitable for highlighting spatial relationships and depth, commonly used to enhance visual guidance and a sense of depth.
- Inverted shot: The image is filmed upside down, challenging the audience's visual habits, often used to represent confusion or an unstable mental state.

Notes

- Each Static Shot Description in the shot scripts should correspond sequentially to each segment of the story, providing detailed descriptions of characters' expressions, actions, and states.

Prompt for Dataset Construction (Part 3)

Notes

- Do not include any characters in the Setting Description. Ignore the characters in the story and only describe the environmental setting.
- Do not introduce items or backgrounds in the Static Shot Description that are not mentioned in the story. Ensure that the location of actions aligns with the story.
- Each sentence in the plot has context. Use this context to determine the best shot design.
- Ensure that the transitions between different shots in the shot scripts follow creative requirements.
- Approach the creation of shot content from a visual perspective. Consider the best way to present the script visually.
- Only provide content that meets the output format requirements; do not include any explanations.
- When dividing the original plot content, rewrite it into a format suitable for presentation.
- Strictly follow the format and requirements of the output example when writing.

A Sample output

```
{
  "Shot 1": {
    "Plot Correspondence": "Fern is performing a magical ritual at the bell tower to
        awaken an ancient power, which is crucial for her team's quest to protect the
        town from impending danger.",
    "Setting Description": "Twilight, stone street leading to the bell tower, bell
        tower with arches, town streets, distant bell tower with arches, romantic
        atmosphere, mysterious atmosphere, glowing arrow on the ground, fallen leaves,
         pigeons perched on rooftops, soft golden light",
    "Shot Perspective Design": "Medium shot, eye level shot",
    "Onstage characters": ["Fern"],
    "Static Shot Description": "Fern is pressed against the bell tower pillar, with an
         expectant and serious expression. Her right index and middle fingers are
        together, suspended in mid-air, and fine golden particles are pouring out from
         her fingertips."
  },
  "Characters": {
    "Fern": "A beautiful girl with long purple hair and purple eyes, wearing a silver
        butterfly hair accessory, a black coat and black boots, and a white dress
        under the coat. She holds a wooden staff in her hand.",
    "Frieren": "A young white-haired female elf with twin ponytails, blue-green eyes,
        pointed elf ears, a pair of red earrings, a white wizard robe with gold trim
        and brown boots, holding a staff with a round ruby on the top.",
    "Himmel": "A young boy with short blue hair and blue eyes, wearing a blue knight
        uniform and a white cape, holding a long sword in his hand, with a gentle yet
        firm expression."
  }
}
```

prompt alignment metrics, such as scene-level consistency and Character Interaction, while the overall alignment score remains within a narrow error margin.

These findings validate the effectiveness of ViStoryBench-Lite as a representative subset of the full benchmark. This is particularly important for settings where large-scale human involvement (e.g. user study) or commercial API evaluation is required, such as user studies or commercial platform assessments. By using ViStoryBench-Lite, researchers and practitioners can achieve efficient and cost-effective evaluation without compromising result reliability.

E.2. Full Evaluation on ViStoryBench-Lite

We present the results of several selected methods on the ViStoryBench-Lite benchmark in the main paper. To facilitate a more comprehensive comparison, we provide the complete evaluation results for all the methods on the lite


TheaterGen
StoryGen (Mixture)
StoryGen (Auto-Regressive)
StoryAdapter (Img-ref scale5)
StoryAdapter (Img-ref scale0)
UNO
StoryDiffusion (Text-only ver.)
StoryDiffusion (Img-ref ver.)
GPT4o
Gemini
Copy-Paste Baseline


Albrm
Doubao
MOKI
Morphic Studio
Typemovie
AnimDirector (SD3 ver.)
MovieAgent (ROICtrl ver.)
MovieAgent (SD3 ver.)
Vlogger
MMStoryAgent

Figure S4. **Qualitative Result on Story 09**. From left to right are shot1 to shot5. Reference images of each shot's onstage characters is shown in Copy-Paste baseline results.

set, as shown in Table S4. These results serve as a thorough reference for assessing the performance of different approaches from multiple perspectives. Additionally, to ensure reproducibility, we provide the complete prompt alignment evaluation results using **Qwen3-VL-8B-Instruct**, deployed via vLLM offline with a fixed seed of 42, as shown in Table S5.

To further interpret the results, we visualize the normalized performance of each evaluated SOTA method across all twelve dimensions using polar coordinates in Table S6. Each radius is normalized relative to the best score achieved among all models for the corresponding dimension.

Performance Analysis of Various Methods. On ViStoryBench-Lite, we observe a wide spectrum of performance among the evaluated methods.

As evidenced by the quantitative results in the tables, **GPT-4o** [15] and **Gemini-2.5** [5] demonstrate superior prompt alignment capabilities, which can be attributed to their fine-grained comprehension abilities rooted in their LLM foundations. Meanwhile, **Sora2** [24] achieves the best balance across the four key metrics of character consistency (cross and self) and style consistency (cross and self), likely benefiting from its extensive training on visually coherent multi-shot data. The performance gap between earlier methods (e.g., **StoryGen** [18], **TheaterGen** [2], **Vlogger** [45]) and more recent approaches (e.g., **UNO** [36], **OmniGen2** [35]) reflects the natural progression and collective advancement within the research community.

Models such as **Storydiffusion** [42] and **StoryAdapter** [20] exhibit strong prompt alignment, while maintaining balanced performance in generation quality and character similarity. **MovieAgent (SD-3)** [34] and **AnimDirector** [17] achieve consistently high scores across most dimensions, with particularly notable strengths in aesthetics and diversity, indicating their advantages in generating visually appealing and varied outputs. Interestingly, **SEED-Story** [39], which focuses on story continuation tasks, exhibits excellent self-consistency in style but shows weaker performance in other metrics. This observation highlights a potential trade-off between maintaining visual fidelity and introducing meaningful variation in generated content.

Among the commercial and proprietary systems, we observe competitive and often leading performance. **AIbrm** [19] and **Doubao** [1], two commercial Chinese plat-

Figure S5. **Qualitative Result on Story 01**. From left to right are shot1 to shot5. Reference images of each shot's onstage characters is shown in Copy-Paste baseline results.

forms, show highly stable performance across almost all metrics. **Doubao**, in particular, demonstrates outstanding results in both prompt alignment and generation quality, with a strong balance between character rendering fidelity and stylistic coherence. These results suggest that closed-source commercial tools have made significant strides in multi-shot storytelling, although their internal pipelines remain opaque.

Overall, these results highlight the multidimensional nature of high-quality story visualization and suggest that no single model excels uniformly across all criteria. These quantitative findings align consistently with our qualitative visual observations and common sense, thereby validating the credibility and reliability of our proposed evaluation metrics for story visualization tasks.

Table S4. **Quantitative Results of Various Story Visualization Methods on ViStoryBench-Lite**. Results highlighted with a gray background are excluded from ranking, for example, SEED-Story is trained on only three animations and does not aim for generalization, while the Copy-Paste Baseline directly pastes the character reference image as the output. For certain methods, we evaluate multiple inference configurations and report all corresponding results. indicate the first, second, third, fourth, and fifth performance, respectively. **CSD**: Style Similarity; **CIDS**: Character Similarity; **PA**: Prompt Alignment Score (CI: Character Interaction, IA: Individual Action); **CM**: OCCM; **Inc**: Inception Score; **Aes**: Aesthetics Score; **CP**: Copy-Paste. I: With image reference; T: Only text input; A: Auto-regressive mode; superscript ^{k} means scale=k. *Note: **PA** scores are based on Gemini-3 Pro.*

Method	Model	CSD↑		CIDS↑		PA↑					CM↑	Inc↑	Aes↑	CP
		Cross	Self	Cross	Self	Scene	Shot	CI	IA	Avg.				
Copy-Paste Baseline	-	0.735	0.770	0.911	0.993	0.42	1.60	0.76	0.85	0.91	92.76	5.46	4.39	0.550
Story Image Method														
StoryGen [18] A	SD1.5	0.405	0.562	0.405	0.591	0.59	1.78	0.58	0.55	0.88	52.98	7.15	4.09	0.277
StoryGen [18] I	SD1.5	0.396	0.551	0.396	0.602	0.59	1.94	0.55	0.25	0.83	52.53	7.67	4.09	0.224
StoryGen [18] A I	SD1.5	0.316	0.617	0.316	0.610	0.46	2.02	0.52	0.30	0.82	40.13	6.25	3.86	0.240
TheaterGen [2]	SD1.5	0.221	0.411	0.354	0.537	1.99	1.92	0.49	0.40	1.20	54.93	13.60	4.94	0.204
StoryDiffusion [42] T	SDXL	0.293	0.680	0.409	0.641	2.61	2.02	1.25	1.05	1.73	67.07	12.99	5.83	0.186
StoryDiffusion [42] I	SDXL	0.409	0.611	0.460	0.575	1.35	2.68	1.34	1.25	1.66	62.48	8.18	5.21	0.251
SEED-Story [39]	SDXL	0.258	0.763	0.559	0.656	1.44	1.67	0.32	0.16	0.90	64.33	4.90	3.81	0.306
Story-Adapter [20] I^0	SD1.5	0.518	0.609	0.490	0.605	1.42	2.53	1.75	1.45	1.79	70.34	11.49	4.89	0.250
Story-Adapter [20] I^5	SD1.5	0.371	0.758	0.425	0.619	1.35	2.40	1.46	1.40	1.65	61.39	12.03	4.80	0.217
Story-Adapter [20] T^0	SD1.5	0.343	0.515	0.430	0.547	1.48	2.67	1.76	1.60	1.88	65.32	12.72	5.12	0.203
Story-Adapter [20] T^5	SD1.5	0.353	0.752	0.416	0.634	1.31	2.44	1.40	1.30	1.61	61.57	10.59	4.85	0.220
UNO [36]	FLUX1	0.425	0.648	0.512	0.630	3.12	2.25	1.98	1.75	2.28	70.88	10.50	5.13	0.287
OmniGen2 [35]	DiT	0.491	0.648	0.576	0.668	3.18	2.38	2.15	1.90	2.40	73.44	8.21	5.21	0.298
CharaConsist [33]	FLUX1	0.333	0.646	0.347	0.539	3.24	2.37	1.82	1.63	2.27	62.10	10.84	5.78	0.216
QwenImageEdit-2509 [28]	DiT	0.404	0.614	0.482	0.541	3.37	2.10	2.45	2.20	2.53	61.27	10.56	5.46	0.249
Story Video Method														
Vlogger [45] T	SD1.4	0.240	0.462	0.369	0.524	1.12	2.25	1.60	1.35	1.58	77.13	8.41	4.24	0.209
Vlogger [45] I	SD1.4	0.299	0.497	0.373	0.519	1.14	2.24	1.53	1.35	1.57	79.14	8.83	4.24	0.211
AnimDirector [17]	SD3	0.305	0.558	0.423	0.593	3.27	2.11	2.44	2.05	2.47	72.03	9.94	5.60	0.206
MMStoryAgent [37]	SDXL	0.261	0.661	0.385	0.598	2.44	1.94	1.14	0.95	1.62	58.24	8.09	5.91	0.189
MovieAgent [34]	SD1.5	0.236	0.564	0.372	0.568	0.93	1.81	0.88	0.85	1.12	64.41	10.06	4.69	0.261
MovieAgent [34]	SD3	0.346	0.539	0.433	0.582	3.09	2.28	2.45	1.90	2.43	67.29	12.04	5.32	0.199
Commercial Platform														
MOKI [22]	-	0.214	0.694	0.372	0.621	2.29	1.56	0.58	0.45	1.22	45.96	10.36	**5.79**	0.211
MorphicStudio [23]	-	**0.577**	0.628	0.603	0.677	3.01	2.31	1.77	1.32	2.10	60.79	9.00	4.96	0.234
AIbrm [19]	-	0.412	**0.730**	0.557	0.740	3.06	2.18	1.67	1.55	2.12	75.53	9.53	5.72	0.223
ShenBi [21]	-	0.275	0.575	0.418	0.585	3.49	2.35	2.65	2.11	2.65	61.33	11.60	5.07	0.197
Typemovie [32]	-	0.325	0.646	0.464	0.621	2.34	2.17	1.62	1.40	1.88	74.14	11.15	5.32	0.168
Doubao [1]	-	0.367	0.695	0.446	0.642	3.88	2.41	3.23	2.65	3.04	65.23	9.88	5.61	0.255
Multi-modal Large Model (Language, Image and Video)														
GPT-4o* [15]	-	0.481	0.680	0.420	0.522	3.82	2.82	3.58	3.12	3.34	69.33	9.02	5.49	0.209
Gemini-2.0* [6]	-	0.361	0.573	0.573	0.677	3.26	2.46	2.43	2.00	2.54	74.82	10.12	4.91	0.266
Gemini-2.5* [11]	-	0.447	0.657	0.553	0.642	3.89	2.32	3.26	2.90	3.09	64.86	10.54	5.61	0.255
Gemini-3.0 Pro* [12]	-	0.385	0.622	0.581	0.653	3.94	2.58	3.71	3.25	3.37	59.97	**12.50**	5.54	0.244
Seedream-4.0 [10]	-	0.369	0.585	0.280	0.539	3.82	2.59	3.20	2.68	3.07	49.45	12.12	5.21	0.201
Sora2* [24] I	-	0.515	0.713	**0.766**	**0.839**	3.08	2.76	2.91	2.50	2.81	**81.42**	6.53	4.72	0.158
Sora2* [24] T	-	0.365	0.685	0.364	0.561	3.01	2.83	2.96	2.33	2.78	72.67	9.68	4.52	0.158

F. Method Evaluation Detail on ViStoryBench

In this section, we report how we adapt each method to the ViStoryBench test. In general, we strive to implement reasonable inputs for reference character images and shot script prompts on each work as much as possible. We make efforts to standardize inputs, such as adjusting output resolution to a 16:9 aspect ratio whenever feasible. To guarantee reproducibility, we fix the random seed. Additionally, we implement mechanisms for inputting reference images and adapting lengthy shot script prompts. These adaptations enable methods to generate continuous image results of stories.

Table S5. **Qwen-Based Prompt Alignment Evaluation of Story Visualization Methods on ViStoryBench and ViStoryBench-Lite**. While the Copy-Paste Baseline directly pastes the character reference image as the output. For certain methods, we evaluate multiple inference configurations and report all corresponding results. indicate the first, second, third, fourth, and fifth performance, respectively. **PA**: Prompt Alignment Score (CI: Character Interaction, IA: Individual Action); I: With image reference; T: Only text input; A: Auto-regressive mode; superscript ^{k} means scale=k.

Method	Model	PA (lite)↑					PA (full)↑				
		Scene	Shot	CI	IA	Avg.	Scene	Shot	CI	IA	Avg.
Copy-Paste Baseline	-	0.30	2.29	0.80	1.11	1.12	0.30	2.29	0.80	1.11	1.12
Story Image Method											
StoryGen [18] A	SD1.5	0.66	2.75	0.85	1.12	1.35	0.65	2.62	0.84	1.10	1.30
StoryGen [18] I	SD1.5	0.71	2.57	0.82	0.98	1.27	0.72	2.65	0.85	1.07	1.32
StoryGen [18] A I	SD1.5	0.48	2.73	0.74	1.18	1.28	0.56	2.68	0.83	1.14	1.30
TheaterGen [2]	SD1.5	2.28	2.32	0.55	0.73	1.47	2.16	2.26	0.49	0.67	1.39
StoryDiffusion [42] T	SDXL	2.73	3.52	1.23	1.38	2.21	2.77	3.48	1.40	1.45	2.27
StoryDiffusion [42] I	SDXL	1.40	3.43	1.62	1.82	2.07	1.38	3.35	1.73	1.72	2.04
SEED-Story [39]	SDXL	1.62	2.79	0.32	0.44	1.29	1.64	2.57	0.54	0.51	1.32
Story-Adapter [20] I^{0}	SD1.5	1.60	3.48	2.00	2.05	2.28	1.68	3.47	2.17	2.07	2.35
Story-Adapter [20] I^{5}	SD1.5	1.65	3.48	1.83	1.74	2.17	1.67	3.50	1.99	1.86	2.26
Story-Adapter [20] T^{0}	SD1.5	1.81	3.54	2.35	2.12	2.45	1.83	3.51	2.28	2.11	2.43
Story-Adapter [20] T^{5}	SD1.5	1.57	3.46	1.84	1.77	2.16	1.64	3.49	1.99	1.87	2.25
UNO [36]	FLUX1	3.11	3.46	2.38	1.94	2.72	3.03	3.53	2.22	1.88	2.67
OmniGen2 [35]	DiT	3.16	3.56	2.43	2.03	2.80	3.09	3.59	2.48	2.07	2.81
CharaConsist [33]	FLUX1	2.54	3.59	1.56	1.40	2.27	2.50	3.53	1.61	1.39	2.26
QwenImageEdit-2509 [28]	DiT	3.24	3.55	2.64	2.22	2.91	2.98	3.53	2.46	2.07	2.76
Story Video Method											
Vlogger [45] T	SD1.4	1.41	3.31	2.12	1.95	2.20	1.36	3.27	2.10	1.88	2.15
Vlogger [45] I	SD1.4	1.37	3.28	1.96	1.81	2.10	1.35	3.24	2.07	1.90	2.14
AnimDirector [17]	SD3	3.37	3.59	2.98	2.17	3.03	3.32	3.61	2.97	2.28	3.04
MMStoryAgent [37]	SDXL	2.64	3.19	1.10	1.25	2.04	2.55	3.30	1.34	1.24	2.11
MovieAgent [34]	SD1.5	0.96	3.07	0.86	0.96	1.46	0.95	3.10	0.80	0.79	1.41
MovieAgent [34]	SD3	3.09	3.59	2.90	2.31	2.97	3.14	3.58	2.96	2.32	3.00
Commercial Platform											
MOKI [22]	-	2.49	2.32	0.54	0.60	1.49	-	-	-	-	-
MorphicStudio [23]	-	3.00	3.47	2.12	1.88	2.62	-	-	-	-	-
AIbrm [19]	-	2.92	3.43	1.73	1.64	2.43	-	-	-	-	-
ShenBi [21]	-	3.48	3.51	3.11	2.25	3.09	-	-	-	-	-
Typemovie [32]	-	2.34	3.51	1.94	1.82	2.40	-	-	-	-	-
Doubao [1]	-	3.80	3.60	3.51	2.73	3.41	-	-	-	-	-
Multi-modal Large Model (Language, Image and Video)											
GPT-4o* [15]	-	3.68	3.79	3.55	2.95	3.49	-	-	-	-	-
Gemini-2.0* [6]	-	3.08	3.55	2.68	2.07	2.84	-	-	-	-	-
Gemini-2.5* [11]	-	3.61	3.48	3.19	2.72	3.25	-	-	-	-	-
Seedream-4.0 [10]	-	3.61	3.58	3.23	2.75	3.29	-	-	-	-	-
Sora2* [24] I	-	3.05	3.56	3.14	2.20	2.99	-	-	-	-	-
Sora2* [24] T	-	2.83	3.62	2.92	2.21	2.89	-	-	-	-	-

Figure S6. **Sample Results of Copy-Paste Baseline**.

Table S6. **Best Normalized Polar Visualization of SOTA Methods on ViStoryBench-Lite**. Anti-Clockwisely, **Character Similarity**: Self-similarity, Cross-similarity; **Style Similarity**: Self-similarity, Cross-similarity; **Prompt Alignment**: Scene, Shot, Character Interaction (CI), Individual Action (IA), Average Score; **Generation Quality**: Aesthetics; **Diversity**: Inception Score.

F.1. Copy-Paste Baseline

The Copy-Paste Baseline aims to verify the correctness of certain metrics by constructing the simplest possible generation method, which involves directly copying and pasting characters into images. For example, this baseline demonstrates that the metrics in ViStoryBench can perform an accurate calculation in areas such as CIDS, CSD, Copy-paste

rate, and Matched Character Count calculation.

We obtain the image for the current shot by simply stitching together the images of the onstage characters in each shot, resulting in the image outcome for the current shot. A sample is shown in Figure S6.

F.2. Story Image Methods

F.2.1. StoryDiffusion

The original StoryDiffusion [42] is primarily designed for image generation using multiple reference images of different characters and multi-shot prompts, where each prompt includes a character name and description. We implement the following adaptations for ViStoryBench evaluation:

① Since the native implementation can only handle short texts that cannot exceed the model's maximum sequence length limit (77 tokens), similarly to StoryGen adaption, we employ the grouped encoder sd_embed [43] to address this issue.

② Given StoryDiffusion does not support inputting reference images of multiple characters while generating one shot, we sort characters based on their appearance frequency in the full shot script, from highest to lowest, and only introduced the highest-priority characters in one shot.

③ StoryDiffusion predefines various style templates, each containing detailed style descriptions, including quality words. We insert the character prompt into the specified position within the style description, i.e., between the style word and the quality word, to achieve richer style expression.

④ During testing, we obtain two types of results: those based on image reference (img ref) and those based solely on text (text only). We report the metrics for both types of results separately.

F.2.2. Story-Adapter

The original Story-Adapter [20] generates multiple images through multiple prompts and does not inherently support input image references. We implement the following adaptations for ViStoryBench evaluation:

① Incorporating Reference Images: Story-Adapter performs multiple rounds of iterative inference. In the first round, it generates all the storyboard images, which are then used as image references for regeneration in subsequent rounds. This multi-round process enhances both prompt alignment and style consistency of the images. We leverage this inherent image-referencing capability of Story-Adapter by inputting the image of the current onstage character into the pipeline during the first round of generation, thereby achieving image generation with reference images.

② We test the results of different iteration rounds (scale 0 and scale 5) under two modes: text-only and image-Ref, and reported the results of all four methods.

F.2.3. UNO

UNO [36]'s original implementation supports various reasoning methods such as one2one, one2many, many2one, and many2many. Among these, the many2many approach involves inputting multiple images along with prompts that provide simple descriptions of the images. We exclusively evaluate the many2many method, which is capable of completing the ViStoryBench task. For each shot, we generate the corresponding image result by inputting the shot's prompt and the images of the onstage characters.

F.2.4. OmniGen2

As a general-purpose image generation model, OmniGen2 is evaluated for its visual consistency capability. Scene, plot, action, and shot design descriptions are concatenated into a prompt. The first reference image of every character appearing in the shot is supplied as a visual reference. Each shot is generated individually. The previous shot's image is excluded from the context, as empirical tests indicated performance degradation when including such references.

F.2.5. CharaConsist

CharacterConsist enables training-free, cross-shot character identity "locking". We evaluate its performance on key metrics such as CIDS_Self.

① "Setting Description" serves as the background prompt, "Character Description" as the foreground prompt, and "Static Shot Description" + "Shot Perspective Design" as the action prompt. Since CharacterConsist does not support explicit character references, no reference images are injected.

② The most frequently appearing character in each story is selected as the protagonist, with "prompt_en" used as the unified foreground character descriptor across all shots.

③ The first shot is generated as the identity reference. Subsequent frames reuse the "id_fg_mask" and "id_bg_mask" from the first frame. Each generation first performs a pre-run ("is_pre_run"=True) before formal generation. Spatial parameters ("spatial_kwargs") are propagated throughout to ensure character and background consistency across shots.

F.2.6. QwenImageEdit-2509

This model is specifically designed for multi-image consistency and is evaluated for its visual coherence capability. Scene, plot, action, and shot design descriptions are concatenated as the prompt. The first reference image of each character in the shot is provided as a visual reference. Each shot is generated sequentially. The image from the previous shot is omitted from the input, as tests demonstrated that including it led to output degradation.

F.2.7. StoryGen

The original StoryGen [18] supports the integration of previously generated image-text pairs as a context to con-

struct image sequences that align with the input shot scripts. We implement the following adaptations for ViStoryBench evaluation:

① Mix Inference Method: The mix method behaves similarly to the multi-image-condition method for the first frame, while subsequent frame generation follows the auto-regressive method.
② When image references are absent (e.g., no onstage characters), the first shot uses aggregated story-wide character descriptions and images for consistent initial representation. For subsequent shots, a dynamic sliding window blends historical generation results as input, maintaining temporal coherence and mitigating quality degradation from long-range dependencies.
③ Grouped Encoder for Ultra Long Prompts: To handle ultra-long prompts and prevent information loss due to truncation in CLIP models, we employ a grouped encoder sd_embed [43] to address this issue.
④ Resolution Adjustment: We modify the resolution settings to approximate a 9:16 aspect ratio, specifically 512×912, to align with the generation resolutions of other methods.
⑤ Deterministic Random Seed Strategy: To enhance test controllability, we adopt a deterministic random seed strategy.

F.2.8. TheaterGen

The original TheaterGen [2] only supports text input, where an LLM is used to parse the description of each character and the overall image of the text. Subsequently, each character's independent image is generated through the IP-Adapter, and then these images are placed in their corresponding positions using detection and segmentation models. Finally, the final image is generated under the guidance of the character images.

Since the method does not open-source the code for invoking the LLM part, we supplement the corresponding code to primarily obtain the bounding box coordinates of each IP image required for model input. Afterward, following TheaterGen's setup, we place the character reference images in our dataset into the corresponding positions to generate final results.

F.2.9. SEED-Story

The original SEED-Story [39] focus on generating long multi-modal stories and their visualization through an auto-regressive approach. SEED-Story requires the user to provide an initial image and text description, then proceeds with the story generation, using each output as the next input. Due to the story-continuation nature, it is different from other approaches (SEED-Story was only trained on the StoryStream dataset [39], which contains data from only three cartoon series, the model does not have generalization).

Input Handling: We only adapt our data in the visualization stage, and there is no need for the previous story generation. For the first input image (000start), select the first reference image of the first character. In the pre-visualization input data, add the prompt words of this role to the beginning of the prompt list to ensure the first shot's prompts are not compromised.

F.3. Story Video Method

F.3.1. MovieAgent

Similar to Vlogger, MovieAgent [34] also receives a story and utilizes an LLM to generate a series of fine-grained descriptions. Additionally, ROIctrl includes the bounding boxes of characters. We directly map shot prompts onto Vlogger's fine-grained descriptions for generation. We only perform the step from prompt to image, without further generating a video, and use this image as the final result.

F.3.2. AnimDirector

AnimDirector [17] does not support input reference images. It utilizes LLM to supplement character and scene descriptions based on simple prompts provided by users, refining them into complete story sequences by scenes. Visual images are generated through the Stable Diffusion 3 [7] model, filtered by VLM, and subsequently used to generate videos. We adopt the following strategies to complete the test on ViStoryBench:

① We directly input the prompt of each shot from ViStoryBench as the story sequence prompt to generate images. After filtering by VLM, we obtain the results without proceeding with the subsequent video generation steps.
② To accommodate ultra-long prompts, we employ grouped encoder sd_embed [43] to address this issue.
③ We modify the original resolution of 1024×1024 to a resolution close to 16:9, specifically 768×1344.
④ We fix the random seed to ensure reproducibility.

F.3.3. MMStoryAgent

MMStoryAgent [37] does not support the input of reference images. We only utilize the Image Agent mentioned in the code to generate images from prompts, without further generating a video. As MMStoryAgent is based on StoryDiffusion, we also employ grouped encoder sd_embed [43] to adapt to long prompts.

F.3.4. Vlogger

The original Vlogger [45] takes a short story and utilizes an LLM to generate a series of fine-grained descriptions. These descriptions encompass character and object descriptions, video script descriptions (in both Chinese and English), characters and objects present, duration settings, etc. We directly map shot prompts onto Vlogger's fine-grained descriptions, selecting the first frame of each shot from the generated video as the result. Similar to StoryDiffusion,

vlogger does not support inputting reference images of multiple characters while generating one shot. We adopt a similar solution to that of StoryDiffusion.

F.4. Commercial Software

Given the absence of API or similar call methods for the following commercial softwares and their intricate interaction processes, we employ external annotators to generate image results for all the commercial software mentioned below. The annotators adhered to a predefined collection instruction and protocol, and all generated results subsequently underwent a rigorous quality check by the authors. All the following methods were tested between May 1 and May 7, 2025.

F.4.1. MOKI

When generating shots on MOKI [22], it is necessary to select one of the provided painting style options. To best replicate the effects of real human usage, we instruct the annotators to choose the option that most closely aligns with the style of the character reference images for each script.

Given that MOKI restricts the maximum number of reference characters in a story to three, we sort the characters based on their appearance frequency among all shots. The three characters with the highest frequency were then added as reference characters to maximize the performance of the model.

We generate images of MOKI using Chinese version dataset. MOKI has a limit on the length of each shot's prompt, capped at 60 Chinese characters. Consequently, we make certain omissions in the prompt for each shot. For instance, we only used the Static Shot Description and Plot Correspondence sections as input. If the character count still exceeded the limit, we employ an LLM to abbreviate the text while preserving key information.

For each shot, the platform generated four images at a time. We select the first image as the final result.

F.4.2. MorphicStudio

MorphicStudio [23] restricts each shot to include only one reference character. Consequently, we rank the frequency of character appearances in the script and use this order as a priority to select the current onstage reference character for each shot.

MorphicStudio allows uploading 2 to 10 images for the same reference character. For characters with only one reference image, we upload an additional identical image to meet the minimum requirement of two images. For characters with multiple reference images, we upload all available reference images. Therefore, we do not provide the Copy-Paste Rate for MorphicStudio because all reference images were uploaded and used for generation, making it incalculable.

For each shot, the platform generates four images at a time. We select the first image as the final result.

F.4.3. AIbrm

AIbrm [19] restricts each shot to a maximum of two reference characters. We employ the same sorting methodology as utilized in MorphicStudio.

The character creation process in AIbrm involves uploading a real human image, selecting a style from the provided options, and inputting a character prompt to generate the character. For the real human category in our dataset, we upload images and chose the "realistic" style. For virtual human and non-human categories, since image uploads were not feasible, we select the closest available style. Subsequently, we input the character prompts from our dataset across all styles to complete the character creation.

F.4.4. ShenBi

Firstly, when creating a project, ShenBi [21] requires selecting a generation style from a list. We choose the style that is closest to the reference images in the dataset.

When creating characters, we upload the character images along with their corresponding prompts to generate new character images, which are subsequently utilized within the method to produce shot images.

ShenBi restricts each shot to a maximum of three reference characters. We employ the same sorting methodology as utilized in MorphicStudio.

The output of ShenBi is in the form of videos. We extract the first frame of each scene video as the final result.

F.4.5. Typemovie

Firstly, when creating a project, Typemovie [32] requires selecting a generation style from a list. We choose the style that is closest to the reference images in the dataset.

The character creation process in Typemovie involves uploading a real human image, selecting a style from the provided options, and inputting a character prompt to generate the character. For the real human category in our dataset, we simply upload reference images. For virtual human and non-human categories, since image uploads were not feasible, we select the closest available style. Subsequently, we input the character prompts from our dataset across all styles to complete the character creation.

Typemovie restricts each shot to include only one reference character. We employ the same sorting methodology as utilized in MorphicStudio.

F.4.6. Doubao

We conduct our tests using the grayscale test version of the "Image Generation" model on the Doubao homepage [1], dated April 27, 2025.

Since this image generation model only supports uploading a single image, we employ a sorting method similar to

that used in MorphicStudio to prioritize characters and selected the highest-priority character to upload as a single reference image.

The prompt used during generation was (translated to English): "This is an image of the protagonist <character_name>. Next, please generate storyboard scenes based on the protagonist's image and the script I provide. The script for the first scene is as follows: <shot_prompt>." We perform multiple rounds of generation in one session to obtain the desired consistent image results for one story.

F.5. Multi-modal Large Models

To integrate Multi-modal Large Models into the ViStoryBench evaluation framework, we adopt the following key adaptation strategies:

① **Atomic Shot Processing:** Each "shot" defined within ViStoryBench was treated as an independent generation request to the model, ensuring focused processing for individual narrative segments.

② **Comprehensive Prompt Engineering:** For every shot, a structured textual prompt was meticulously crafted. This prompt amalgamated all critical textual information provided by ViStoryBench, including plot details, scene descriptions, character portrayals, camera perspective guidelines, and the desired aspect ratio for the output image.

③ **Direct Visual Referencing:** Character reference images, after undergoing a standardization pre-processing pipeline (e.g., resizing, color space conversion), were directly incorporated as visual inputs for each shot's generation request. This aimed to guide the model in rendering characters consistent with their specified appearances.

④ **Conversational Context Continuation:** To foster narrative coherence across sequential shots, the model's inherent capability to process conversational history was leveraged. The most recent interaction cycles, encompassing the prompt for the preceding shot and the model's response (which includes the generated image), served as contextual information for the subsequent shot's generation task.

F.5.1. GPT-4o

OpenAI's GPT-4o [15] represents a general-purpose multimodal understanding and generation model. We evaluate whether its prompt alignment capability directly translates to high-quality visual storytelling.

Scene, plot, action, and shot design descriptions are concatenated into a unified prompt. The image from the previous shot and the first reference image of all characters appearing in the current shot are used as visual references. Each shot is generated sequentially.

F.5.2. Gemini-2.0

Google's Gemini-2.0 [6] represents a general-purpose multimodal understanding and generation model. We evaluate whether its prompt alignment capability directly translates to high-quality visual storytelling.

Scene, plot, action, and shot design descriptions are concatenated into a unified prompt. The image from the previous shot and the first reference image of all characters appearing in the current shot are used as visual references. Each shot is generated sequentially.

F.5.3. Gemini-2.5 (NanoBanana)

Gemini-2.5 [5] represents a general-purpose multimodal understanding and generation model. We evaluate whether its prompt alignment capability directly translates to high-quality visual storytelling.

Scene, plot, action, and shot design descriptions are concatenated into a unified prompt. The image from the previous shot and the first reference image of all characters appearing in the current shot are used as visual references. Each shot is generated sequentially.

F.5.4. Seedream-4.0

Seedream-4.0 [10] represents a multimodal Image understanding and generation model. Scene, plot, action, and shot design descriptions are concatenated into a unified prompt. The image from the previous shot and the first reference image of all characters appearing in the current shot are used as visual references. Each shot is generated sequentially.

F.5.5. Sora2

Sora2 [24] represents a long-video generation model with native multi-shot capability. We aim to evaluate its long-range narrative comprehension and multi-shot visual consistency.

① Scene, plot, action, and shot design descriptions for each shot are concatenated to form the input prompt.

② In "img_ref" mode, the first reference image of every character in the story is provided to define overall visual style and character baselines. In text_only mode, image references are disabled.

③ TransNetV2 [31] is used for shot boundary detection and keyframe extraction, followed by manual verification to match keyframes to their corresponding shots.

④ Stories that fail to generate due to copyright restrictions or realistic portrait style limitations are excluded; only successfully generated stories are included in average metric calculations.

G. Details of User Study

The user study component of our research was time-limited, concluding in mid-May 2025 due to project constraints. Its main purpose was to validate our automated evaluation metrics. As the user feedback showed a strong correlation with

Scoring Criterial and Interface of User Study

Character Identification Consistency: Based on the provided story visualization results, please assess the character id consistency of characters throughout the story and provide a score.
Scoring Criteria:

- 0: There is a lack of fundamental ID consistency, with nearly every image featuring different characters, indicating an almost complete absence of images with matching characters.
- 1: In a smaller subset of images (about 10-30%), the main characters demonstrate mutual consistency.
- 2: In a moderate number of images (around 30-60%), the main characters can be treated as having mutual consistency.
- 3: In a substantial subset of images (approximately 60-80%), the main characters exhibit mutual consistency. However, a minor portion of images still shows inconsistencies in character representation.
- 4: The main characters are consistently identifiable across the vast majority of images.

Environment Consistency: Based on the provided story visualization results, please assess the environment consistency throughout the story and provide a score.
Scoring Criteria:

- 0: There is a lack of fundamental environmental consistency; under the same environmental description, the generated scenes exhibit neither consistent style nor content.
- 1: At a glance, the scenes appear to have some level of consistency, such as similar styles. However, upon closer inspection, the content is entirely different and lacks any coherence.
- 2: There is a certain level of consistency in the style and semantic information of the image scenes, such as the presence of similarly styled beds and windows. However, inconsistencies exist in either the style or specific content, for instance, while tables and desk lamps are present in both, the desk lamps themselves are not similar.
- 3: The majority of image have consistent semantic information, with style and specific content being largely uniform.
- 4: Nearly all image scenes exhibit strong consistency in both style and specific content, akin to the effect of video recording within the same scene over a continuous timeframe.

Subjective Aesthetics Score: Based on the provided results, please assess the aesthetics of the story and provide a score.
Scoring Criteria:

- 0: Most characters have very obvious generation problems, such as distorted faces, extra/missing limbs, or the painting style is very uncomfortable for humans to watch. Or the image quality is extremely poor.
- 1: Characters have obvious generation problems, such as extra/missing limbs, distortion, etc., but there is no discomforting content. Or the image quality is poor.
- 2: Over 80% of the characters have no obvious physical problems, and there is no obvious content that causes physical discomfort, but the visual experience is poor. Almost all the content of the images, such as character poses, is completely the same, lacking variation.
- 3: Over 80% of the characters have no obvious physical problems. Mediocre picture books with ordinary visual experience, lacking variation and storytelling in images.
- 4: Over 80% of the characters have no obvious physical problems. Excellent and beautiful picture books that can be commercialized, with rich content, beautiful details, diversity, and interest, and obvious storytelling.

Table S7. **Results of User Study**. For certain methods, we evaluate multiple inference configurations and report all corresponding results. indicate the first, second, third, fourth, and fifth performance, respectively. I: With image reference; T: Only text input; A: Auto-regressive mode; superscript ^{k} means scale=k.

Method	Model	Character Identification Consistency ↑	Environment Consistency ↑	Subjective Aesthetics ↑
Story Image Method				
StoryGen [18] A	SD1.5	*0.10*	*0.12*	*0.05*
StoryGen [18] I	SD1.5	0.18	0.27	0.13
StoryGen [18] A I	SD1.5	0.37	0.22	0.10
TheaterGen [2]	SD1.5	0.35	0.55	0.30
StoryDiffusion [42] T	SDXL	2.72	2.73	2.45
StoryDiffusion [42] I	SDXL	2.62	2.33	2.30
SEED-Story [39]	SDXL	2.05	2.11	1.05
Story-Adapter [20] I^0	SD1.5	2.55	2.62	2.80
Story-Adapter [20] I^5	SD1.5	2.90	2.98	2.68
Story-Adapter [20] T^0	SD1.5	2.33	2.23	2.50
Story-Adapter [20] T^5	SD1.5	3.10	2.67	2.68
UNO [36]	FLUX1	3.20	**3.10**	3.02
Story Video Method				
Vlogger [45] T	SD1.4	0.87	1.07	0.67
Vlogger [45] I	SD1.4	1.30	1.33	1.08
AnimDirector [17]	SD3	2.52	2.28	1.77
MMStoryAgent [37]	SDXL	2.27	2.78	2.55
MovieAgent [34]	SD1.5	1.90	1.95	1.55
MovieAgent [34]	SD3	2.45	2.47	1.83
Commercial Platform				
MOKI [22]	-	1.73	2.20	2.55
MorphicStudio [23]	-	2.60	2.53	2.39
AIbrm [19]	-	3.42	2.97	3.15
ShenBi [21]	-	2.74	2.89	2.48
Typemovie [32]	-	2.25	2.35	2.00
Doubao [1]	-	**3.63**	3.02	3.25
Multi-modal Large Model (Language, Image and Video)				
GPT-4o* [15]	-	3.08	3.06	**3.28**
Gemini-2.0* [6]	-	2.84	2.84	2.26

the automated scores, we confirmed the reliability of the automated method. Therefore, subsequent model evaluations were predominantly carried out using this approach.

In correlation analysis, we exclude results from StoryGen [18], as its limited generation quality caused human evaluators to penalize character consistency scores, This is due to the generated character deviating significantly from the distribution of typical characters, leading to mismatched human expectations.

To comprehensively assess the visual quality and consistency of story visualization results, we conducted a structured human evaluation on the ViStoryBench-Lite benchmark. This benchmark contains a diverse subset of stories across multiple genres and character settings, making it suitable for evaluating both identity consistency and visual storytelling fidelity.

Evaluation Dimensions. Each story visualization result was evaluated on three key dimensions:

- **Character Identification Consistency:** Measures whether the main characters remain visually consistent and recognizable across different shots in the story.
- **Environment Consistency:** Assesses the consistency of environmental elements—such as furniture, architecture, or background settings—across the sequence of generated images.
- **Subjective Aesthetics Score:** Evaluates the overall visual appeal of the story, including character quality, composition, artistic style, and storytelling clarity.

Each dimension was scored on a 5-point Likert scale (from 0 to 4), based on clear qualitative criteria provided to the annotators as shown in the scoring interface below. For example, a score of 0 in Character Consistency indicates a complete lack of identifiable characters across shots,

whereas a score of 4 indicates nearly perfect character continuity.

Annotation Interface and Process. We developed a web-based annotation interface below that displays:

- The full set of generated images for a story.
- Relevant textual annotations, such as the story prompt and shot descriptions.
- A structured table for scoring each of the three criteria.

The interface was designed to facilitate efficient and focused annotation, allowing annotators to toggle between image sequences and text prompts while assigning scores.

Annotator Pool and Assignment Strategy. We recruited **20 human annotators** with prior experience in visual content evaluation, including graduate students and crowd workers trained on our scoring rubric. Given the large number of models, stories, and dimensions to evaluate, we adopted a **balanced partial assignment** strategy: each annotator was assigned only a subset of the full evaluation set, but we ensured that:

- **Every model-story pair** received at least **10 independent ratings per dimension**.
- Annotators were randomly assigned different stories and methods to avoid bias.
- Each task contained only a manageable number of stories (typically 6–8), to reduce fatigue.

This design helped scale the annotation process while maintaining evaluation reliability.

Aggregation and Analysis. For each model and story, we aggregated the scores by taking the mean across annotators. We also report standard deviation across annotators to reflect inter-rater variability. The collected annotations form the human evaluation benchmark for comparing model performance on ViStoryBench-Lite in terms of character coherence, environmental stability, and visual storytelling quality.

H. Details of Prompt Alignment Evaluation

To evaluate how well the generated images align with the input shot prompts, we employ GPT-4.1 [25] as an automatic evaluator. The LLM is prompted to assign a Likert-scale rating (0–4) for each image-prompt pair across several semantic dimensions, including scene correctness, camera composition, and character actions. The average of these subtask scores yields the final Alignment Score used in Table S4.

To better analyze the capabilities and limitations of each method, we further break down the Alignment Score into four interpretable sub-scores: **Scene Score**, **Shot Score**, **Character Interaction**, and **Individual Action**, with the final score computed as an equally weighted average of these components. Each dimension focus on a specific aspect of visual-textual alignment. For example, the Scene Score evaluates the match between background or scene attributes and the prompt, while the Camera Score assesses adherence to cinematographic framing (e.g., close-up, long shot). The action scores reflect whether the actions of characters (either collectively or individually) are faithful to the described narrative.

From the results, we observe that recent video generation methods adapted for story visualization, such as **AnimDirector** and **MovieAgent (SD3)**, achieve the highest alignment scores across most categories. AnimDirector leads in **Scene Score** (3.61) and **Character Interaction** (3.24), while MovieAgent (SD3) excels in **Individual Action** (2.50). In contrast, conventional image-generation methods like **StoryGen** or **SEED-Story** generally perform poorly, with significantly lower scores across all metrics.

This detailed analysis reveals that high-quality story visualization requires coherent handling of both low-level visual elements (like camera and scene) and high-level semantics (like character intent and interaction), underscoring the necessity of specialized multi-modal reasoning for prompt alignment.

H.1. Character Interaction

To assess fine-grained alignment between visual content and textual prompts—particularly focusing on interactions between characters—we introduce a semantic consistency evaluation protocol targeting *Character Interaction*. This task evaluates whether the generated image accurately captures the described relational dynamics between two or more characters, such as hugging, fighting, handing over an object, or sitting together. Such interactions are crucial for evaluating story-level coherence and the model's ability to capture nuanced inter-character behavior.

H.2. Shooting Method (Shot)

To assess the framing and compositional accuracy of generated images from a cinematic perspective, we propose a *shot-type alignment* evaluation. This task examines how well the image conforms to the specified camera distance (e.g., close-up, medium shot, wide shot) and camera angle (e.g., eye-level, high-angle, low-angle) provided in the textual prompt. Accurate shot framing is essential for conveying narrative focus, emotional tone, and spatial arrangement—hallmarks of professional visual storytelling.

H.3. Static Shot Description (Scene)

To evaluate the fidelity of background and environment rendering, we introduce a *static shot grounding* task. Unlike the character-centric evaluations, this task focus on

non-character elements—including environmental context, background objects, spatial layout, and overall ambient mood. It measures whether these visual elements align semantically with the scene descriptions in the prompt, such as "a classroom with a blackboard and wooden desks" or "a cozy bedroom with warm lighting and starry wallpaper." This evaluation is critical for assessing the model's holistic scene understanding.

H.4. Individual Action

To further probe character-level grounding and behavioral fidelity, we propose an *individual action consistency* evaluation. This task isolates the action of a specific named character described in the prompt—such as "Tom is raising his right hand"—and checks whether the generated image faithfully represents this behavior. We crop the character from the image using a detected bounding box and perform feature-based comparisons for evaluation. This task offers a focused lens on the model's ability to correctly associate and render discrete physical actions with the correct character identity.

H.5. Effectiveness of Prompt Alignment Metric

To ensure the evaluation protocol accurately captures the nuanced alignment between generated visualizations and textual prompts, we conducted an in-depth case study analyzing prompt alignment across a diverse set of generated shots. As illustrated in Figure S7, we dissected the prompt into multiple semantically rich components—such as plot correspondence, setting, camera perspective, character presence, and static shot descriptions—and compared how different evaluation methods respond to these elements.

For each shot, we collected expert annotations to indicate the relevance and completeness of the generated content with respect to these prompt components. We then compared three types of automatic scoring signals: CLIP-based similarity, GPT-based scoring from VinaBench [9], and the proposed ViStoryBench evaluation protocol. The case study revealed that while CLIP and VinaBench scores often misalign with semantic and stylistic fidelity, our ViStoryBench protocol consistently correlates better with expert judgments, particularly in terms of narrative consistency and cinematographic correctness (as highlighted by the green checkmarks).

This iterative analysis informed the design of ViStoryBench's evaluation prompts and scoring rubric, enabling a more reliable and interpretable assessment of visual story generation quality.

H.6. Stability of VLM-based Automatic Evaluation

A common concern with evaluations that leverage Large Language Models (LLMs), such as GPT-4.1 [25], is the potential for variability in their outputs, which could affect the reliability of the results. To rigorously validate the stability and robustness of our automated evaluation framework for prompt alignment, we conducted a detailed stability analysis.

Experimental Setup for Stability Test We performed 3 to 5 independent evaluation runs for each prompt-adherence metric (*Scene Score*, *Character Interaction*, *Individual Action*, and *Camera Score*). This analysis was conducted on a representative subset of our benchmark, comprising 5 diverse stories (01, 09, 27, 41, 53) and the results from three key methods: UNO, StoryDiffusion, and Story-Adapter. By repeatedly evaluating the same set of generated images, we can precisely quantify the variance attributable to the LLM evaluator itself.

Results and Analysis The results demonstrate exceptionally low variance across all evaluation runs. The standard deviation for each metric was consistently an order of magnitude smaller than the performance gaps observed between different models in our main experiments. For instance, after aggregating the scores across the selected methods and stories, we observed the following mean scores and standard deviations:

- *Scene Score*: 2.82 ± 0.03
- *Character Interaction*: 2.57 ± 0.04
- *Individual Action*: 2.40 ± 0.08
- *Camera Score*: 3.23 ± 0.03

These minimal standard deviations confirm that the random fluctuations in the GPT-4.1 [25] evaluator are negligible. This high level of consistency ensures that the performance differences we report in our benchmark are meaningful reflections of model capabilities, rather than artifacts of evaluation noise. This finding strongly supports the reliability of using our VLM-based protocol for large-scale, automated story visualization assessment.

H.7. Correlation Analysis of Qwen-base and GPT-based Evaluation

Based on the Prompt Alignment scoring results from the ViStoryBench-Lite dataset, we conducted a correlation analysis between the scores provided by **GPT-4.1** [25] and **Qwen3-VL-8B-Instruct** [38]. This analysis aimed to reveal the consistency between these two VLMs in evaluating story visualization methods. We focused on the Prompt Alignment Score (PA) average (Avg.) as it represents the comprehensive indicator of overall performance. The PA score includes sub-items: Scene (scene description), Shot (shot perspective), CI (character interaction), IA (individual action), and Avg. (average). Below is a detailed analysis report.

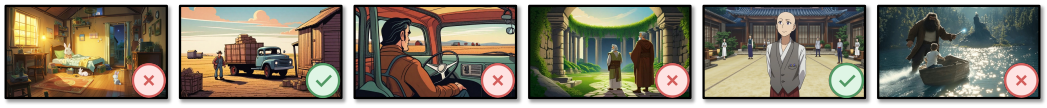
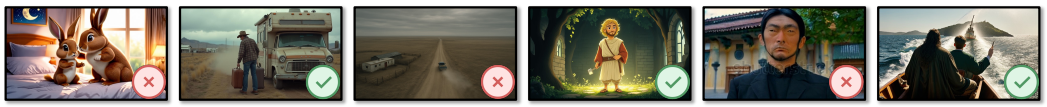
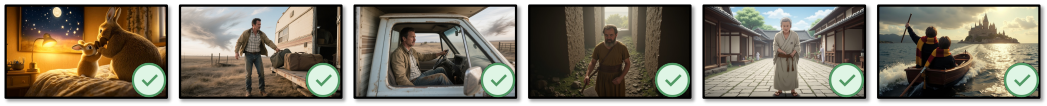
Plot Correspondence	It's time for the little rabbit to go to bed, but he tightly holds onto the big rabbit's ears and refuses to let go.	Ennis knows he must move today because the farm has been sold.	Carrying his memories of Jack, Ennis embarks on a new journey.	Daedalus, commissioned by King Minos, built an incredibly complex labyrinth.	The old man became wealthy because of Princess Kaguya.	They boarded a small boat and left the island. Hagrid waved his wand, and the boat shot forward like an arrow, the splashing waves glittering like crystals in the sunlight.
Setting Description	Nighttime, the little rabbit's bedroom, cozy atmosphere, soft bedding, a small night light on the bedside table, a starry sky painting on the wall, the moon and stars visible outside the window, soft yellow lighting	Early morning, outside an old trailer, surrounded by a desolate farm, wind howling, Ennis stands by the trailer, preparing to pack his belongings.	Early morning, outside an old trailer, surrounded by a desolate farm, wind howling, Ennis sits in the truck's driver seat, ready to depart.	Daytime, inside the labyrinth, gloomy atmosphere, stone walls covered with ivy, moss growing on the ground. Flickering light streams through narrow windows, faint echoes resonate from the depths of the labyrinth, dim lighting	Daytime, Sanuki no Takamuro's new residence, a spacious courtyard, opulent house, bustling servants, bright lighting	Daytime, at sea, the small boat speeding forward, waves splashing, sunlight shining on the water, the outline of the island in the distance, bright lighting.
Shot Perspective Design	Medium shot, eye level shot	Full shot, low-angle shot	Medium shot, eye level shot	Medium shot, high-angle shot	Full shot, high-angle shot	Wide shot, rear view, bird's eye view shot
Characters Appearing	Little Brown Rabbit、Big Brown Rabbit	Ennis Del Mar	Ennis Del Mar	Daedalus	Sanuki no Takamuro	Harry Potter, Hagrid
Static Shot Description	The little brown rabbit sits on the bed, tightly holding the big brown rabbit's ears, with a playful and cute expression. The big brown rabbit sits by the bed, looking gentle, slightly lowering his head to look at the little rabbit	Ennis stands by the trailer, looking resolute yet helpless, loading his belongings onto the truck, the background showing the desolate farm and howling wind.	Ennis sits in the truck's driver seat, looking resolute yet sorrowful, gripping the steering wheel, gazing into the distance through the windshield, the background showing the desolate farm and howling wind.	Daedalus stands at the center of the labyrinth, his expression focused, holding design blueprints in his hand as he surveys the surrounding walls	Sanuki no Takamuro stands in the center of the courtyard, looking around with a contented expression, his hands clasped behind his back.	Harry and Hagrid sit in the boat. Hagrid waves his wand, and the boat speeds forward. Harry grips the side tightly.
High CLIP Score						
High GPT Score (VinaBench)						
⭐ **High GPT Score (ViStoryBench)**						

Figure S7. **Case Study of Prompt Alignment.** We compare the alignment between generated visualizations and the detailed prompt components—including plot correspondence, setting, shot perspective, characters, and static shot description—across six different shots. Each row provides expert annotations on prompt elements, while the bottom three rows indicate whether the generated image achieves a high CLIP score, a high GPT score according to VinaBench, or a high GPT score under the proposed ViStoryBench evaluation. ViStoryBench consistently shows better alignment with visual-semantic fidelity and shot design, as reflected by the green checks.

We extracted the PA Avg. scores (GPT-based and Qwen-based) for each method from Table S4 and Table S5, forming a paired dataset with scores ranging from 0 to 4 points. Statistical correlation analysis was performed, calculating the **Pearson correlation coefficient** (measuring linear correlation) and the **Spearman rank correlation coefficient** (measuring monotonic correlation based on rankings). The calculations and analyses are as follows:

Pearson Correlation Analysis for PA Average. The Pearson correlation coefficient is calculated using the formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}},$$

where X represents GPT-based scores and Y represents Qwen-based scores.

Based on 34 data points, the Pearson correlation coefficient is $r \approx 0.93$. This value close to 1 indicates a strong positive linear correlation between GPT-based and Qwen-based PA Avg. ratings. That is, when GPT scores are higher, Qwen scores tend to be higher as well, and vice versa.

Spearman Correlation Analysis for PA Average. The Spearman correlation coefficient is calculated using the formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where d_i is the difference in ranks, and n is the sample size.

After ranking the scores (higher scores receive higher ranks) and computing the rank differences, the Spearman correlation coefficient is $\rho \approx 0.89$. This value also close to 1 indicates high consistency in the ranking order between the two ratings, demonstrating significant monotonic correlation. This implies that both models similarly evaluate the relative performance of different methods.

Correlation Analysis for PA Sub-items. For comprehensiveness, we also calculated the correlation coefficients for the PA sub-items (Scene, Shot, CI, IA). The average Pearson correlation coefficient ranges from approximately 0.85 to 0.90, and the average Spearman coefficient ranges from 0.82 to 0.88, both indicating moderate to strong correlations.

Conclusion of Correlation Analysis. Through correlation analysis of the Prompt Alignment scores on the ViStoryBench-Lite dataset, we found that the ratings from GPT-4.1 and Qwen3-VL are highly correlated (Pearson $r \approx 0.93$, Spearman $\rho \approx 0.89$). Both GPT-4.1 and Qwen3-VL are advanced VLMs with powerful multimodal understanding capabilities. They are trained on similar datasets, resulting in similar evaluation criteria for prompt alignment. The strong correlation indicates that in story visualization tasks, both models perform consistently when evaluating prompt alignment capability, demonstrating interchangeability between them.

I. Details of Character Identification Similarity

I.1. Calculation

Figure **??** illustrates the computation pipeline of our Character Identification Similarity (CIDS) metric. CIDS quantifies both *self-similarity* (within generated images) and *cross-similarity* (between generated and reference images) through a four-stage computational pipeline:

① **Character Detection.** Grounding DINO localizes character regions using text prompts.
   ◦ *Reference images*: Crops character regions with edge trimming (minor boundary adjustments).
   ◦ *Generated images*: May fail to detect characters (returns empty result), indicating identity inconsistency.

② **Feature Extraction.** Extracts 512D embeddings from cropped regions:
   ◦ *Realistic characters*: ArcFace/AdaFace/FaceNet tri-model ensemble (robust facial features).
   ◦ *Stylized characters*: CLIP ViT-L/14 (semantic alignment).

③ **Bipartite Matching.** Solves optimal character correspondence via Hungarian algorithm:
   ◦ Computes cosine similarity matrix between reference/generated features.
   ◦ Matches characters maximizing global similarity (excludes failed detections).

④ **Scoring.** Final metric: CIDS $= \frac{1}{N}\sum_{i=1}^{N} \cos(\mathbf{v}_{\text{ref}}^{(i)}, \mathbf{v}_{\text{gen}}^{(i)})$ where N = number of matched pairs, $\mathbf{v}$ = feature vectors.

Table S8. **Cross-CIDS Metric with Different Reference Image**. "**Dataset Reference**" refers to results calculated with reference images in ViStoryBench dataset, "**Generated Reference**" refers to results calculated with generated reference images of methods. All results below are obtained on ViStoryBench-Lite.

Method	Dataset Reference	Generated Reference
MOKI [22]	0.292	0.338
AIbrm [19]	0.559	0.683
ShenBi [21]	0.347	0.389

I.2. Impact of Reference Image Selection on Cross-CIDS Metric

In certain methods, the character reference images or features used for generation are not directly sourced from our dataset but are instead synthesized through an additional generation stage. A common scenario involves converting real-person reference images from our dataset into stylized versions—such as anime-style characters—resulting in visual appearances that may differ substantially from the original subjects. These discrepancies in reference image selection can significantly impact the Cross-CIDS metric. In the main results tables, we report scores based on the original reference images provided in the dataset. For a more comprehensive comparison, we additionally report Cross-CIDS scores computed using the synthesized reference characters, as shown in Table S8.

J. Details of Style Similarity Calculation

Figure S8 illustrates the computation pipeline of our Style Similarity metric. Adapted from CSD [30, 44], this metric captures both *self-similarity* (within generated images) and *cross-similarity* (between generated and reference images) by analyzing style-specific features extracted via CSD-CLIP [30].

The computation consists of three key steps:

① Each image is encoded into visual embeddings using a CLIP [29] vision encoder pre-trained on large-scale style datasets;

② The extracted features are passed through CSD layers to disentangle content and style representations, retaining only the style components;

③ Pairwise cosine similarity is computed between the resulting style embeddings to measure stylistic alignment.

This design enables fine-grained comparison of artistic and stylistic consistency, independent of content semantics.

K. Details of OCCM

Onstage Character Count Matching (OCCM) metric relies on an upstream character detector to obtain the detected character count (D). Consequently, the accuracy of the

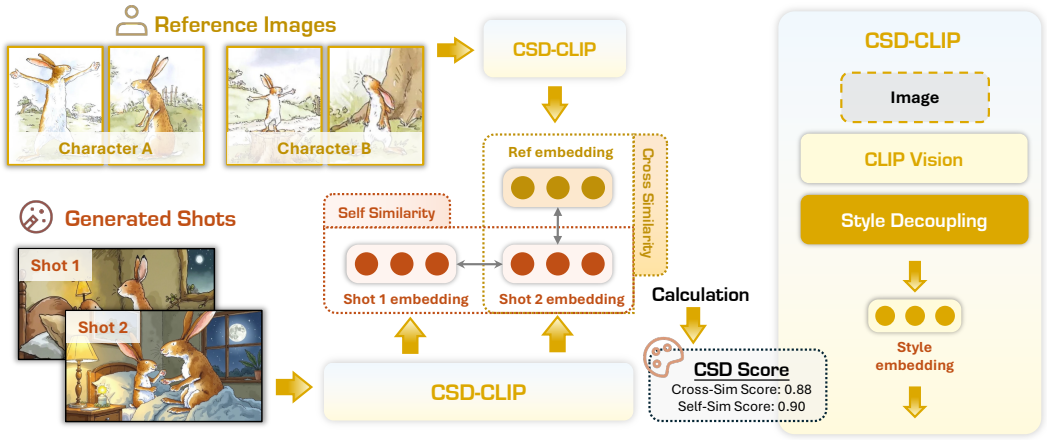


Figure S8. **Style Similarity Calculation Pipeline**. Evaluating both cross-similarity and self-consistency by computing cosine similarity between the style features of two images.

OCCM score is inherently bounded by the detector's performance, which may be affected by factors like heavy occlusion or extreme artistic styles. We opted for expert models as the upstream detector. While expert models can misjudge, VLMs perform worse in hallucination counting. The design of OCCM formula is based on two core principles:

1. **Scale Normalization:** By dividing the absolute error $|D - E|$ by the expected count E, we convert the error into a relative percentage. This allows the metric to fairly evaluate scenes of varying scales, ensuring that the same relative error receives a similar penalty, whether in a single-character scene ($E = 1$) or a multi-character scene ($E = 10$).
2. **Non-linear Penalty:** We employ an exponential function to penalize the relative error. Unlike a linear penalty, the exponential function leads to a gentle score decay for small errors but a sharp drop for larger ones. This characteristic better aligns with human perception: missing one out of ten characters is a minor flaw, whereas missing five constitutes a critical failure.

L. Details of Copy-Paste Detection

To rigorously evaluate whether the generated image is merely a replication of a specific reference image (denoted as the anchor or target reference $\mathbf{r}_0$) rather than a generalized synthesis from the provided character concept, we employ a Softmax-based Copy-Paste Score.

Let $\mathbf{g}$ be the unit-normalized feature vector of the generated image, and $\mathcal{R} = \{\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_N\}$ be the set of unit-normalized feature vectors for the input reference images, where $\mathbf{r}_0$ represents the primary reference subject to copy-paste detection, and $\{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ serves as the set of auxiliary references for the same character.

We first calculate the cosine similarity between the generated image and each reference image in the set $\mathcal{R}$. To quantify the exclusivity of the match between $\mathbf{g}$ and the target $\mathbf{r}_0$ relative to other references, we formulate the score as a probability distribution using a temperature-scaled Softmax function:

$$\text{CopyRate}(\mathbf{g}|\mathcal{R}) = \frac{\exp(\mathbf{g}^\top \mathbf{r}_0/\tau)}{\sum_{k=0}^{N} \exp(\mathbf{g}^\top \mathbf{r}_k/\tau)} \tag{S1}$$

where τ is a temperature hyperparameter set to 0.01. This low temperature value sharpens the distribution, making the metric highly sensitive to the nearest neighbor in the feature space.

The resulting **Copy-Paste Rate** ranges from 0 to 1. A score approaching 1 indicates that the generated image $\mathbf{g}$ is significantly more similar to the specific reference $\mathbf{r}_0$ than to any other provided references, suggesting a "copy-paste" overfitting behavior. Conversely, a lower score implies that the generated features are either distributed among multiple references or have successfully generalized beyond the specific appearance of $\mathbf{r}_0$. The final metric is averaged across all generated samples for each character.

M. Benchmark Evaluation Efficiency

Table S9. **Computational Efficiency of Evaluation Metrics**.

Metric	Scope	Time	Notes
Aesthetics Score	Single image	0.026s	per generated image
Style Similarity	Pair images	0.046s	cross or self
Character Similarity	Pair images	0.450s	cross or self
Inception Score	Total data	8.057s	full dataset
Prompt Alignment*	Single image	25.173s	per generated image

* Longest computation due to LLM inference constraints

To ensure the reproducibility and transparency of our computational experiments, we provide detailed information regarding the hardware setup and evaluation runtime. All experiments are conducted using high-performance GPU accelerators (e.g., NVIDIA H800 with 80GB memory), and we ensure consistent measurement conditions across metrics.

Table S9 reports the average computational cost associated with each evaluation metric. These metrics differ significantly in scope and computational complexity. For example, aesthetic scoring is performed on individual images and is highly efficient, averaging only 0.026 seconds per image. Style and character similarity metrics, which operate on image pairs, are slightly more demanding, especially character similarity (0.450 seconds per pair), likely due to the use of deep feature extractors.

In contrast, metrics such as the Inception Score are computed over the entire dataset and thus have a higher runtime (8.057 seconds), though only once per dataset. The most computationally intensive metric is the Prompt Alignment score, requiring over 25 seconds per image. This is primarily due to the involvement of large language model (LLM) inference, which introduces latency constraints.

Prompt for Character Interactions

Task Definition

You will be provided with an image and a text prompt describing the main character's action. As an experienced evaluator, your task is to evaluate the semantic consistency between the image and the text prompt, according to the scoring criteria. This evaluation focus specifically on whether the action of the main character in the image aligns with the action described in the text.

Scoring Criteria

When evaluating the semantic consistency between an image and its corresponding text prompt, the following aspects are crucial:

- **Relevance**: Does the image show the main character performing the action or behavior mentioned in the text? The action in the image should match the core description provided in the text.
- **Accuracy**: Does the image depict the action correctly according to the text prompt? Any specific details related to the action, such as gestures, posture, or environment, should align with the description.
- **Completeness**: Does the image show the main character completing the entire action as described in the text? The image should not omit important parts of the action or behavior.

Scoring Range

Based on these criteria, you will assign a score from 0 to 4 that reflects the degree of semantic consistency between the image and the text prompt:

- **Very Poor (0)**: No correlation. The image does not reflect any aspect of the action described in the text prompt.
- **Poor (1)**: Weak correlation. The image addresses the text in a very general way but misses most details and accuracy of the action.
- **Fair (2)**: Moderate correlation. The image depicts the action to some extent, but there are several inaccuracies or missing details.
- **Good (3)**: Strong correlation. The image accurately portrays most elements of the action with minor inaccuracies or omissions.
- **Excellent (4)**: Near-perfect correlation. The image closely aligns with the text prompt and portrays the main character's action with high accuracy and precision.

Input format

Every time you will receive a text prompt and an image.
Please carefully review the image and text prompt. Before giving a score, please provide a brief analysis of the above evaluation criteria, which should be very concise and accurate.

Output Format

Analysis: <Your analysis>
Score: <Your Score>

Prompt for Shooting Evaluation

Task Definition

You will be provided with an image and a text prompt describing the shot type of the image. As an experienced evaluator, your task is to assess whether the generated image meets the specified shot requirements based on the evaluation criteria.

Additional Material

Instruction: You are a professor evaluator. Below is information about different shot types and shot distances. Please evaluate whether the generated image meets the requested shot type.

Shot Distance Descriptions

- **Long Shot**: Shows the relationship between characters and their environment, typically used to display the scene or environment.
- **Full Shot**: Shows the full body of a character, commonly used to display movement or the full scene.
- **Medium Long Shot**: Starts from above the character's knees, capturing part of the environment.
- **Medium Shot**: Captures the character from the waist up.
- **Close-Up**: Captures the character from the chest up.
- **Extreme Close-Up**: focus on the character's head or face, with the background and environment typically blurred or not visible.

Angle Descriptions

- **Eye Level Shot**: The camera is positioned at the subject's eye level.
- **Low Angle Shot**: The camera is positioned below eye level, shooting upward, emphasizing the character's power or size.
- **High Angle Shot**: The camera is positioned above eye level, shooting downward, often minimizing the subject's significance.
- **Bird's Eye View**: Camera shot taken from directly above, providing an overview of the scene.
- **Tilted Shot**: The camera is intentionally tilted to create a sense of imbalance or tension.
- **Perspective Compression**: A technique that emphasizes depth and the relationship between foreground and background through perspective.

Scoring Range

A score between 0 and 4 will be assigned based on how well the shot type aligns with the content described in the prompt:

- **Very Poor (0)**: The image does not meet any shot or angle requirements.
- **Poor (1)**: The image meets some but not most of the shot or angle requirements.
- **Fair (2)**: The image partially meets the shot or angle requirements, but some elements are off.
- **Good (3)**: The image meets most of the shot or angle requirements.
- **Excellent (4)**: The image fully meets all of the shot and angle requirements.

Input Format

You will receive a text prompt and an image. Please carefully review the image and text prompt. Provide an analysis followed by a score.

Output Format

Analysis: <Your analysis >
Score: <Your score>

Prompt for Static Shot Evaluation

Task Definition

You will be provided with an image and a text prompt that describes the **background**, objects, and mood of the scene (excluding characters). Your task is to evaluate the consistency between the **background and objects** described in the prompt and what is visually represented in the image.

Evaluation Criteria

When assessing the semantic consistency between the image and the text prompt, focus on how well the **background and non-character elements** in the image match the description provided in the text. The evaluation should be based on the following aspects:

- **Relevance**: The image should clearly relate to the primary background elements and objects described in the text. It should reflect the main setting and environment described, without introducing irrelevant or unrelated features.
- **Accuracy**: Check if the specific details mentioned in the text are correctly represented in the image. This includes any mentioned objects, scenery, environmental conditions (e.g., weather, lighting), and relevant background elements.
- **Completeness**: Evaluate whether the image accurately includes all critical background elements described in the text. The image should reflect the key details and setting, not leaving out essential aspects of the described background or scene.
- **Context**: The image should maintain the context of the description. If the text describes a specific environment or atmosphere, the image must capture that context appropriately, considering the described mood and setting elements.

Scoring Criteria

Based on these factors, the image will be assigned a score from **0 to 4**, indicating the degree of consistency between the image and the description in the text:

- **Very Poor (0)**: No correlation. The image completely fails to reflect the background or objects described in the text.
- **Poor (1)**: Weak correlation. The image touches on the background or objects in a very general sense but misses most of the important details or has significant inaccuracies.
- **Fair (2)**: Moderate correlation. The image contains some relevant background and objects, but several important details are missing or inaccurately represented.
- **Good (3)**: Strong correlation. The image accurately represents most of the described background and objects with minor omissions or inaccuracies.
- **Excellent (4)**: Near-perfect correlation. The image perfectly captures the background and objects as described in the text, leaving no significant details missing or inaccurate.

Input Format

You will receive a text prompt and an image. Please carefully review the image and text prompt. Provide an analysis followed by a score.

Output Format

Analysis: <Your analysis>
Score: <Your Score>

Prompt for Individual Action Evaluation

Task Definition

For each evaluation, you will receive a text prompt, an image, and a character name. Your task is to **first extract the individual action or behavior of the specified character from the text prompt, then determine whether the image accurately reflects this description for that character**, and finally assign a score based on the criteria.

Evaluation Process

- **Extract Action Information**: Carefully extract the specific action or behavior described for the given character (character name) from the text prompt.
- **Image Comparison**: Examine the image to determine whether the specified character's action matches the extracted description.
- **Analyze and Score**: Analyze the match according to the scoring criteria and assign a score.

Scoring Criteria

Focus on the following aspects when evaluating:

- **Relevance**: Does the image show the specified character performing the action or behavior described in the text?
- **Accuracy**: Are the details of the character's action in the image (such as posture, gestures, environment) consistent with the text description?
- **Completeness**: Does the image fully depict the character completing the entire action as described, without omitting important parts?

Assign a score from 0 to 4 based on the degree of semantic consistency:

- **0 (Very Poor)**: No correlation. The image does not reflect any aspect of the described action.
- **1 (Poor)**: Weak correlation. The image only generally addresses the text, missing most details and accuracy.
- **2 (Fair)**: Moderate correlation. The image depicts the action to some extent but with several inaccuracies or missing details.
- **3 (Good)**: Strong correlation. The image accurately portrays most elements of the action, with only minor inaccuracies or omissions.
- **4 (Excellent)**: Near-perfect correlation. The image closely aligns with the text prompt and depicts the character's action with high accuracy and completeness.

Output Format

Analysis: <Your analysis>
Score: <Your Score>

References

[1] ByteDance Inc. Doubao ai assistant. https://www.doubao.com/, 2024. Accessed: 2025-04-16. 2, 12, 14, 15, 19, 22

[2] Junhao Cheng, Baiqiao Yin, Kaixin Cai, Minbin Huang, Hanhui Li, Yuxin He, Xi Lu, Yue Li, Yifei Li, Yuhao Cheng, et al. Theatergen: Character management with llm for consistent multi-turn image generation. *arXiv preprint arXiv:2404.18919*, 2024. 2, 12, 14, 15, 18, 22

[3] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 5

[4] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, 2023. 5

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 12, 20

[6] Google DeepMind. Gemini 2.0 flash: Native image generation in google ai studio. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/, 2025. Accessed: 2025-04-16. 3, 14, 15, 20, 22

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 18

[8] Patrick Esser, Sumith Kulal, Ajay Jitkasan, Andkhuja Rakhimov, Dawid Filip, Meng Du, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 3

[9] Silin Gao, Sheryl Mathew, Li Mi, Sepideh Mamooler, Mengjie Zhao, Hiromi Wakaki, Yuki Mitsufuji, Syrielle Montariol, and Antoine Bosselut. Vinabench: Benchmark for faithful and consistent visual narratives. *arXiv preprint arXiv:2503.20871*, 2025. 24

[10] Lixue Gong, Xiaoxia Hou, Fanshi Li, Liang Li, Xiaochen Lian, Fei Liu, Liyang Liu, Wei Liu, Wei Lu, Yichun Shi, et al. Seedream 2.0: A native chinese-english bilingual image generation foundation model. *arXiv preprint arXiv:2503.07703*, 2025. 14, 15, 20

[11] Google. Nano Banana: Image editing in Gemini just got a major upgrade. https://gemini.google.com/nano-banana, 2025. 14, 15

[12] Google. Nano Banana Pro: Image editing in Gemini just got a major upgrade. https://deepmind.google/models/gemini-image/pro/, 2025. 14

[13] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, 2018. 3, 6

[14] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL*, 2016. 3, 6

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 12, 14, 15, 20, 22

[16] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, 2019. 3, 6

[17] Yunxin Li, Haoyuan Shi, Baotian Hu, Longyue Wang, Jiashun Zhu, Jinyi Xu, Zhen Zhao, and Min Zhang. Anim-director: A large multimodal model powered agent for controllable animation video generation. In *SIGGRAPH Asia*, 2024. 12, 14, 15, 18, 22

[18] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *CVPR*, 2024. 3, 6, 12, 14, 15, 17, 22

[19] MagicLight AI. Brmgo: Ai-powered tool for story script generation. https://brmgo.cn/, 2025. Accessed: 2025-04-16. 12, 14, 15, 19, 22, 26

[20] Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization, 2024. 2, 3, 12, 14, 15, 17, 22

[21] Maoyan Entertainment. Shenbi: Ai-powered scriptwriting tool by maoyan. https://shenbi.maoyan.com/, 2025. Accessed: 2025-04-16. 14, 15, 19, 22, 26

[22] Meitu Inc. Moki: Ai short film creation tool. https://www.moki.cn, 2024. 2, 14, 15, 19, 22, 26

[23] Morphic, Inc. Introducing morphic studio. https://www.morphic.com/, 2024. Accessed: 2025-04-16. 2, 14, 15, 19, 22

[24] OpenAI. Sora 2. https://sora.chatgpt.com/, 2025. Accessed: 2025-11-11. 3, 12, 14, 15, 20

[25] OpenAI. Gpt-4.1: Advanced large language model for natural language understanding and generation. https://openai.com/research/gpt-4-1, 2025. Accessed: 2025-04-16. 23, 24

[26] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. Renderme-360: a large digital asset library and benchmarks towards high-fidelity head avatars. *NeurIPS*, 2023. 5

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[28] Qwen Team. Qwen-image technical report. https://qwen.ai/blog/qwen-image, 2025. 14, 15

[29] 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 models from natural language supervision. In *ICML*, 2021. 26

[30] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024. 26

[31] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 20

[32] TypeMovie Team. Typemovie: Text-to-video storytelling with style and rhythm. https://typemovie.art, 2024. Accessed: 2025-04-16. 14, 15, 19, 22

[33] Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. *arXiv preprint arXiv:2507.11533*, 2025. 14, 15

[34] Mike Zheng Shou Weijia Wu, Zeyu Zhu. Automated movie generation via multi-agent cot planning, 2025. 3, 12, 14, 15, 18, 22

[35] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 2, 12, 14, 15

[36] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 12, 14, 15, 17, 22

[37] Xuenan Xu, Jiahao Mei, Chenliang Li, Yuning Wu, Ming Yan, Shaopeng Lai, Ji Zhang, and Mengyue Wu. Mm-storyagent: Immersive narrated storybook video generation with a multi-agent paradigm across text, image and audio. *arXiv preprint arXiv:2503.05242*, 2024. 14, 15, 18, 22

[38] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 24

[39] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 2, 3, 6, 12, 14, 15, 18, 22

[40] Zilyu Ye, Jinxiu Liu, JinJin Cao, Zhiyang Chen, Ziwei Xuan, Mingyuan Zhou, Qi Liu, and Guo-Jun Qi. Openstory: A large-scale open-domain dataset for subject-driven visual storytelling. In *CVPR*, 2024. 3, 6

[41] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 1

[42] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *NeurIPS*, 2024. 3, 12, 14, 15, 17, 22

[43] Shudong Zhu(Andrew Zhu). Long prompt weighted stable diffusion embedding. https://github.com/xhinker/sd_embed, 2024. 17, 18

[44] Cailin Zhuang, Yaoqi Hu, Xuanyang Zhang, Wei Cheng, Jiacheng Bao, Shengqi Liu, Yiying Yang, Xianfang Zeng, Gang Yu, and Ming Li. Styleme3d: Stylization with disentangled priors by multiple encoders on 3d gaussians. *arXiv preprint arXiv:2504.15281*, 2025. 26

[45] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. In *CVPR*, 2024. 3, 12, 14, 15, 18, 22