STORYDYNAMICS: HOLISTIC EVALUATION OF STORY VISUALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A coherent story visualization requires more than generating individually plausible images. Correct entities should appear, recurring characters should remain recognizable, their interactions should reflect the narrative, visual states should be faithfully depicted, those states should change at the appropriate point in the story, and the resulting frames should remain visually faithful to the text. Existing benchmarks capture subsets of these properties, but do not provide a unified evaluation of them jointly. We introduce STORYDYNAMICS, an evaluation framework that assesses story visualization across six complementary dimensions: entity count, character identity, character–object interaction, visual state, state transition, and visual quality. To support this evaluation, we construct a benchmark of 93 human-validated English and Chinese stories spanning diverse narrative lengths, domains, themes, and visual styles. Each story is paired with the reference information needed to evaluate both frame-level properties and how observable states evolve throughout the narrative. Using STORYDYNAMICS, we evaluate 21 general-purpose and story-specific generation models. Our benchmark evaluation shows that strong performance on individual dimensions does not necessarily translate to coherent story visualization, that character-object interaction is the most vulnerable capability across story length and language, and that story-specific models do not consistently achieve holistic gains over general-purpose text-to-image models. Our results highlight the importance of jointly evaluating both static and temporal properties when assessing story visualization models.

1 INTRODUCTION

Recent advances in deep generative models (Podell et al., 2024; Rombach et al., 2022; Goodfellow et al., 2020; Labs et al., 2025) have enabled applications across creative arts (Ko et al., 2023), user interface design (Chen et al., 2025), and video games (Buongiorno et al., 2024). Story visualization has similarly emerged as a compelling application in entertainment and education (Rizvic et al., 2019; Tong et al., 2018). Recent work (Li et al., 2019; Shen & Elhoseiny, 2025; Zhou et al., 2024) demonstrates increasingly capable models for transforming textual narratives into sequences of visual frames.

Yet, visualizing a story involves more than generating a collection of individually plausible images. *Story visualization* aims to generate sequences of images or videos that faithfully depict a textual narrative. Unlike conventional text-to-image generation, stories unfold through interdependent events. Characters and objects should remain recognizable across frames, visual attributes and backgrounds should remain coherent, interactions should reflect the narrative, and later scenes should preserve consequences established earlier in the story. A successful story visualization model should therefore model the narrative as a whole rather than treat each frame independently (see Fig. 1).

Existing benchmarks capture important aspects of story visualizations across different perspectives. Text-to-image benchmarks such as OneIG-Bench (Chang et al., 2026) and TIIF-Bench (Wei et al., 2025) examine generations across input conditions, *e.g.*, prompt length and language, while evaluating properties including image-text alignment and visual quality. Meanwhile, story-specific benchmarks extend generic text-to-image evaluations towards sequential generations. ViStoryBench (Zhuang et al., 2026) and VinaBench (Gao et al., 2025) evaluate properties such as character

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Overview of StoryDynamics. (A)** The benchmark contains 93 stories covering 9 domains, 55 themes, 4 visual styles, multiple languages, and both short and long stories, resulting in 4,740 frames for each evaluated model. **(B)** Example evaluation of a five-frame story. For each frame, StoryDynamics checks whether the correct entities, character identities, interactions, visual states, and image quality are preserved. It tracks how character, object, and background states change across frames to evaluate whether transitions happen at the correct point in the story.

consistency, and plausibility of story visualizations. LogiStory (Meng et al., 2026), on the other hand, studies whether generated story frames preserve causal relationships as the story progresses.

These benchmarks provide valuable insights into story visualization, but their evaluation emphases remain fragmented. Some focus on cross-frame consistency, while others emphasize prompt variation or causal relationships. Consequently, they do not provide a unified view of the capabilities required for coherent story visualization. This motivates a central question: *How can we obtain a holistic view of a story visualization model's capabilities across the diverse properties that constitute a coherent story?*

To answer, we propose StoryDynamics, a holistic benchmark for evaluating story visualization models across narrative conditions and visual fidelities. StoryDynamics evaluates six complementary properties of a coherent visual story, namely: (i) presence of required entities, (ii) consistency of character identities, (iii) correctness of character-object interactions, (iv) accuracy of character, object and background states, (v) accuracy of the transitions between states, and (vi) visual quality and image-text alignment, while spanning variations in story length, complexity, language, style and domain. By fusing these dimensions, StoryDynamics provides a comprehensive view of generative models' capabilities in translating narratives into coherent visual frames.

Our contributions are as follows:

- We establish a unified evaluation framework that assesses cross-frame consistency, image-text alignment, visual fidelity, state transitions, and fine-grained interactions, providing a holistic characterization of story visualization capabilities within a common setting. We further validate the proposed narrative metrics on an award-winning professionally directed film, where their behavior is consistent with the expected properties of a coherent cinematic narrative.
- We introduce StoryDynamics, a story visualization benchmark comprising 93 human-validated stories spanning multiple languages, story lengths, and levels of complexity. Evaluating each model produces a total of 4,740 generated frames, which we assess along six complementary dimensions designed to capture both individual-frame quality and cross-frame narrative coherence.

- We conduct an extensive empirical evaluation of 21 story visualization models, including general-purpose text-to-image, training-free, trained, agentic, and video-based approaches. Our results show that strong performance on individual dimensions does not necessarily translate to coherent story visualization, with character–object interaction emerging as the most vulnerable capability across story length and language.

2 Related Works

Generative and Story Visualization Models. General-purpose text-to-image models primarily target image fidelity and prompt alignment (Z.ai, 2026; OpenAI, 2026; Wu et al., 2026; 2025a; Labs, 2024; 2025; Podell et al., 2024; Esser et al., 2024), while subsequent work improves character consistency across generated images (Wang et al., 2025a; Wu et al., 2025b; Wang et al., 2025b). Story visualization extends beyond character identity by requiring coherence across an evolving sequence. Existing approaches include training-based methods for cross-frame consistency (Li et al., 2019; Maharana et al., 2021; Liu et al., 2024; Yang et al., 2025), training-free methods that condition on reference images (Zhou et al., 2024; Mao et al., 2026; Tewel et al., 2024; Liu et al., 2025), agentic models that decompose story generation into specialized subtasks (Li et al., 2024; Xu et al., 2025; Wu et al., 2025c), and video-generation models that explicitly model temporal progression (Ho et al., 2022; Zhuang et al., 2024).

Conventional Text-to-Image Evaluations. Text-to-image evaluation has progressed from curated prompt suites such as DrawBench (Saharia et al., 2022) and PartiPrompts (Yu et al., 2022) to benchmarks targeting object counting, spatial relations, image–text faithfulness, compositionality, and dense instruction following (Cho et al., 2023; Hu et al., 2023; Bakr et al., 2023; Huang et al., 2023; Ghosh et al., 2023; Hu et al., 2024). More recent benchmarks further consider reasoning, text rendering, style, multilingualism, prompt length, and complex instructions (Chang et al., 2026; Wei et al., 2025). These benchmarks provide increasingly comprehensive evaluation of individual images, but do not explicitly capture cross-frame dependencies such as persistent entity states or whether state changes occur at the correct point in a narrative.

Story Visualization Evaluations. Story visualization benchmarks range from fixed-character datasets such as PororoSV (Li et al., 2019), FlintstonesSV (Gupta et al., 2018), and DiDeMoSV (Maharana et al., 2022), to open-ended settings such as StorySalon (Liu et al., 2024) and OpenStory++ (Ye et al., 2024). More recent benchmarks evaluate structured aspects of story coherence, including character, style, time, and location constraints in VinaBench (Gao et al., 2025), identity and other cross-frame properties in ViStoryBench (Zhuang et al., 2026), and causal narrative reasoning in LogiStory (Meng et al., 2026). These efforts provide complementary views of story visualization, but no single benchmark jointly evaluates entity presence, character identity, character–object interaction, visual state, state transition, and visual quality within a unified framework. StoryDynamics is designed to address this gap.

3 StoryDynamics

3.1 StoryDynamics Evaluations

Evaluation Dimensions. We define the StoryDynamics evaluation dimensions as follows:

(i) ***Entity Count.*** Stories often involve multiple characters and objects, and missing or introducing entities can alter the meaning of a scene. A faithful visualization should therefore contain the expected number of notable characters and objects described at each narrative step.

(ii) ***Character Identity Consistency.*** Recurring characters should remain visually recognizable throughout the story so that readers can reliably distinguish who is participating in each event. This requires preserving identity-defining visual characteristics across frames despite changes in pose, expression, or viewpoint.

(iii) ***Character–Object Interactions.*** Stories are defined not only by which entities are present, but also by how they interact. A frame may contain the correct character and object while still depicting the wrong event if their interaction does not match the narrative. Character–object interaction therefore captures whether the depicted actions and relationships reflect those described in the story.

(iv) ***Visual States.*** Narrative coherence depends on characters, objects, and environments being depicted in the correct condition at each point in the story. An injured character, a broken object, or a changed environment should remain consistent with what has already been established by the narrative. Visual-state evaluation captures this correspondence between the current narrative state and its visual realization.

(v) ***State Transitions.*** Stories evolve over time, making it important not only to depict the correct states but also to change them at the appropriate moment. A model may depict an individual state correctly while introducing it too early, too late, or reverting to an earlier state in subsequent frames. State-transition evaluation captures whether these changes occur at the correct narrative step and persist consistently as the story progresses.

(vi) ***Visual Quality.*** Narrative correctness alone does not guarantee a satisfactory visual story. Generated frames should also exhibit strong perceptual quality and remain faithful to the accompanying text, ensuring that the story is both visually coherent and semantically aligned with the narrative.

Evaluation Setup. Given a story $S = (x_1, \ldots, x_T)$, where x_t denotes the narrative at step t, a story visualization model generates a corresponding sequence of frames $F = (F_1, \ldots, F_T)$. StoryDynamics evaluates each generated story across six complementary dimensions as aforementioned. Unless otherwise stated, metrics are computed over all relevant instances within a story and subsequently aggregated across the benchmark.

Entity Count Accuracy. Entity count accuracy measures whether each generated frame contains the expected number of characters and story-relevant objects. For characters, both ground-truth and predicted entity sets are restricted to the canonical benchmark characters associated with reference images; entities outside this character set are excluded from both sets. Let $E_{t,k}$ and $\hat{E}_{t,k}$ denote the ground-truth and predicted entity sets at step t, where $k \in \{\text{char}, \text{obj}\}$. We compute

$$\text{Cnt}_k = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\Big[|E_{t,k}| = |\hat{E}_{t,k}|\Big], \qquad k \in \{\text{char}, \text{obj}\}.$$

Character Identity Preservation. Character identity preservation measures whether recurring characters maintain a consistent visual identity throughout the generated story. For models that require character-reference conditioning, generated characters are compared against the corresponding reference image. For models without reference-image conditioning, identity consistency is measured across generated frames in which the same character appears. Let N_{id} denote the number of evaluated character comparisons and $N_{\text{id}}^{\text{match}}$ the number that preserve the expected identity. We compute

$$\text{ID} = \frac{N_{\text{id}}^{\text{match}}}{N_{\text{id}}}.$$

Identity matching focuses on identity-defining characteristics while allowing natural variations in pose, expression, viewpoint, and lighting. Accordingly, ID reflects identity preservation under each model's native conditioning setting, rather than a controlled comparison under a single identity-reference protocol.

Character-Object Interaction (COI) Accuracy. Character-object interaction accuracy measures whether the interactions depicted in each frame agree with the narrative. An interaction is represented as $h = (c, r, o)$, where c, r, and o denote the character, action, and object, respectively. Let $N_{\text{coi}}^{\text{gt}}$, $N_{\text{coi}}^{\text{pred}}$, and $N_{\text{coi}}^{\text{match}}$ denote the numbers of ground-truth, predicted, and correctly matched interactions across a story. We compute

$$\text{F1}_{\text{COI}} = \frac{2N_{\text{coi}}^{\text{match}}}{N_{\text{coi}}^{\text{gt}} + N_{\text{coi}}^{\text{pred}}} \qquad \text{EM}_{\text{COI}} = \frac{1}{|\mathcal{T}^+|} \sum_{t \in \mathcal{T}^+} \mathbf{1}\Big[H_t = \hat{H}_t\Big].$$

Here, H_t and $\hat{H}_t$ are the ground-truth and predicted interaction sets, and $\mathcal{T}^+$ denotes frames containing annotated interactions.

Visual State Accuracy. It measures whether the generated frames depict the visual states specified by the story. For each narrative step, we construct the expected observable states and represent them as *entity–property–value* facts, such as $(\text{mug}, \text{condition}, \text{broken})$ or $(\text{background}, \text{weather}, \text{rainy})$.

States are grounded in the story text, except when outfit, held-item, or injury states have not yet been specified at a character's first appearance, in which case they are initialized from the character reference. These states are carried forward and updated only when the narrative introduces a visually observable change.

For each generated frame, we extract predicted state facts and compare them with the corresponding ground-truth states. For each entity type $k \in \{\text{char}, \text{obj}, \text{bg}\}$, let $N^{\text{gt}}_{\text{state},k}$, $N^{\text{pred}}_{\text{state},k}$, and $N^{\text{match}}_{\text{state},k}$ denote the numbers of ground-truth, predicted, and correctly matched state facts. We compute

$$\text{StateF1}_k = \frac{2N^{\text{match}}_{\text{state},k}}{N^{\text{gt}}_{\text{state},k} + N^{\text{pred}}_{\text{state},k}}, \qquad k \in \{\text{char}, \text{obj}, \text{bg}\}.$$

A state fact is counted as correct when its entity, property, and observable value agree with the corresponding ground truth.

State Transition Accuracy. State transition accuracy measures whether observable states change at the correct point in the narrative. A state transition is represented as $\tau = (n, p, v^-, v^+, t)$, where visual property p of entity n changes from v^- to v^+ at narrative step t. The resulting state persists until another narrative event introduces a subsequent change.

For each entity type, let $N^{\text{gt}}_{\text{trans},k}$, $N^{\text{pred}}_{\text{trans},k}$, and $N^{\text{match}}_{\text{trans},k}$ denote the numbers of ground-truth, predicted, and correctly matched transitions. We compute

$$\text{TransF1}_k = \frac{2N^{\text{match}}_{\text{trans},k}}{N^{\text{gt}}_{\text{trans},k} + N^{\text{pred}}_{\text{trans},k}}, \qquad k \in \{\text{char}, \text{obj}, \text{bg}\}.$$

A transition is matched when the entity, visual property, previous value, new value, and transition step agree with the ground truth.

Visual Quality. Visual quality measures text–image alignment and perceptual quality. UnifiedReward (Wang et al., 2025c) assigns each frame–sentence pair a score $U(x_t, F_t) \in [0, 100]$, which is normalized per frame and averaged across the story. A LAION aesthetic predictor (Schuhmann et al., 2022), operating on L2-normalized CLIP ViT-L/14 embeddings (Radford et al., 2021), produces a raw aesthetic score $a(F_t)$. We compute

$$\text{UR} = \frac{1}{T}\sum_{t=1}^{T}\frac{U(x_t, F_t)}{100}, \qquad \text{AES} = \frac{1}{10T}\sum_{t=1}^{T} a(F_t).$$

Overall Score. For visual state and state transition accuracy, we first average across characters, objects, and backgrounds:

$$\text{StateF1} = \frac{1}{3}\sum_{k\in\{\text{char},\text{obj},\text{bg}\}} \text{StateF1}_k, \qquad \text{TransF1} = \frac{1}{3}\sum_{k\in\{\text{char},\text{obj},\text{bg}\}} \text{TransF1}_k.$$

Let the nine summary scores be

$$(m_1, \ldots, m_9) = (\text{Cnt}_{\text{char}}, \text{Cnt}_{\text{obj}}, \text{ID}, \text{F1}_{\text{COI}}, \text{EM}_{\text{COI}}, \text{StateF1}, \text{TransF1}, \text{UR}, \text{AES}).$$

The overall score is computed as the harmonic mean of the nine component scores, encouraging balanced performance across component metrics.

3.2 Validation Against a Professional Cinematic Narrative

To validate whether the proposed StoryDynamics metrics behave consistently with expectations for a professionally produced visual narrative, we evaluate them on the award-winning film *One Battle After Another*. Professional filmmaking involves deliberate control over narrative continuity, character and object consistency, visual composition, and temporal progression. A professionally directed and edited film therefore provides a meaningful reference for examining whether the proposed metrics assign high scores to the narrative properties they are designed to capture.

The validation follows three steps. First, a VLM (`gpt-5.5-2026-04-23`) derives a textual storyline from the film sequence, describing its characters, objects, interactions, visual states, and state

Table 1: **Validation against a professional cinematic narrative.** We derive the storyline and reference annotations from the video, and evaluate the original video frames as predictions. $\mathrm{Overall}_{\mathrm{all}}$ includes all metrics, while $\mathrm{Overall}_{\mathrm{nar}}$ excludes pretrained visual-quality scores.

	Count		ID	COI		State			Transition			Quality		Overall	
Source	$\mathrm{Cnt}_{\mathrm{char}}\uparrow$	$\mathrm{Cnt}_{\mathrm{obj}}\uparrow$	ID $\uparrow$	$\mathrm{F1}_{\mathrm{COI}}\uparrow$	$\mathrm{EM}_{\mathrm{COI}}\uparrow$	Char.	Obj.	Bg.	Char.	Obj.	Bg.	UR $\uparrow$	AES $\uparrow$	$\mathrm{Overall}_{\mathrm{all}}\uparrow$	$\mathrm{Overall}_{\mathrm{nar}}\uparrow$
Real video	1.000	0.667	1.000	0.875	0.800	0.867	0.889	1.000	1.000	–	–	0.442	0.500	0.735	0.877

transitions. Second, this storyline is converted into the same structured reference annotations used by StoryDynamics. Third, the original film frames are treated as predictions and evaluated against these annotations using the same narrative metrics as in the main experiments. This procedure allows the proposed metrics to be examined on a professionally constructed visual narrative outside the generative-model setting.

For the cinematic validation, $\mathrm{Overall}_{\mathrm{all}}$ is computed over all nine summary metrics, while $\mathrm{Overall}_{\mathrm{nar}}$ excludes UR and AES; undefined entity-level transition scores are omitted when computing TransF1. As shown in Table 1, the film achieves an overall narrative score of $\mathrm{Overall}_{\mathrm{nar}} = 0.877$, with consistently high scores across the constituent narrative dimensions. This outcome aligns with the expected characteristics of professionally directed cinema, in which character continuity, object relationships, interactions, and evolving visual states are deliberately maintained across shots. The result therefore provides evidence that quantities measured by StoryDynamics correspond to meaningful aspects of narrative coherence beyond generated story visualizations.

Visual-quality metrics are considered separately, as UR and AES rely on external pretrained evaluators whose behavior on cinematic footage is not validated here. We therefore focus on $\mathrm{Overall}_{\mathrm{nar}}$ and its narrative dimensions; additional results are in Appendix E.

3.3 Benchmark Formulation

Dataset. StoryDynamics comprises 93 English and Chinese stories spanning 5–100 sentences, multiple visual styles, 9 domains, and 55 themes. Each sentence corresponds to one narrative frame, yielding 4,740 frames per evaluated model. The benchmark varies story length, linguistic complexity, language, visual style, and domain; an overview is shown in Fig. 1.

Story Construction. We construct the story corpus through a controlled generation and validation pipeline. For each theme, we first define its semantic scope and assign a narrative configuration specifying the target story length, linguistic complexity, and language. We then generate $K = 3$ candidate stories for each configuration, allowing variation in characters, events, and narrative progression while preserving the assigned narrative requirements. Construction details are provided in Appendix A.2. Each candidate undergoes two stages of validation using the same predefined rubric. An LLM-based validator first filters candidates that do not satisfy the required narrative, structural, and linguistic criteria. Candidates that pass this stage are subsequently reviewed by human annotators using the same criteria. Only stories that satisfy both stages are retained, resulting in 93 human-validated stories. Further details of the generation and validation procedure are provided in Appendix A.3.

Ground-Truth Construction. For each retained story, we construct the reference information required for evaluation. We first derive identity references for recurring characters from their persistent visual characteristics in the narrative, then convert the story into a sequence of structured state graphs

$$G = (G_1, \ldots, G_T), \qquad G_t = (C_t, O_t, B_t),$$

where C_t, O_t, and B_t describe the expected observable states of characters, objects, and the background at narrative step t. These states are tracked throughout the story, allowing the benchmark to capture both the expected visual state at each step and the transitions between states as the narrative progresses. Full details of the ground-truth construction procedure are provided in Appendix A.4.

Table 2: **Comparison of 21 models on StoryDynamics.** We report performance across the six evaluation dimensions and the **Overall** harmonic mean defined in Sec. 3.1. **Bold** and underlined values denote the best and second-best results. Full results are in Appendix D.

Model	Count		ID	COI		State	Trans.	Quality		Overall ↑
	Cnt_{char} ↑	Cnt_{obj} ↑	ID ↑	F1_{COI} ↑	EM_{COI} ↑	StateF1 ↑	TransF1 ↑	UR ↑	AES ↑	
Text-to-Image Models										
Generative Models										
GLM-Image	0.439	0.336	**0.864**	0.089	0.038	0.260	0.270	0.064	0.639	0.131
GPT-Image-1-mini	0.419	0.410	0.713	0.103	0.035	**0.389**	**0.432**	0.101	0.618	0.148
OmniGen2	0.625	0.445	0.511	0.116	**0.057**	0.281	0.311	0.087	0.625	0.173
Qwen-Image-Edit	0.573	0.459	0.392	0.094	0.042	0.370	0.390	0.136	0.620	0.162
Flux.1	0.638	0.266	0.040	0.033	0.011	0.227	0.212	0.053	0.671	0.050
Flux.2	**0.695**	**0.486**	0.767	**0.134**	0.048	0.380	0.379	**0.151**	0.635	**0.193**
Stable Diffusion XL (SDXL)	0.453	0.218	0.017	0.020	0.010	0.172	0.183	0.046	0.637	0.035
Stable Diffusion 3.5	0.607	0.277	0.039	0.035	0.015	0.235	0.222	0.065	0.651	0.059
Training-Free Models										
CharaConsist	0.447	0.223	0.026	0.037	0.018	0.223	0.198	0.045	**0.688**	0.055
Trained Models										
UNO	0.557	0.264	0.355	0.047	0.021	0.194	0.185	0.044	0.641	0.080
MS-Diffusion	0.380	0.254	0.001	0.003	0.000	0.134	0.127	0.051	0.619	0.000
Story Visualization Models										
Training-Free Models										
StoryDiffusion	0.468	0.180	0.010	0.017	0.014	0.176	0.149	0.033	0.652	0.032
Story Iter	0.392	0.175	0.002	0.004	0.002	0.137	0.122	0.032	0.601	0.007
ConsiStory	0.450	0.186	0.001	0.001	0.000	0.131	0.099	0.015	0.575	0.001
One-prompt-one-story	0.366	0.248	0.002	0.003	0.000	0.149	0.136	0.042	0.588	0.002
Trained Models										
StoryGen	0.380	0.199	0.023	0.001	0.001	0.127	0.114	0.012	0.503	0.003
SEED-Story	0.286	0.163	0.000	0.000	0.000	0.109	0.104	0.006	0.504	0.000
Agentic Models										
Anim-Director	0.642	0.253	0.002	0.011	0.007	0.230	0.181	0.062	0.636	0.012
MMStoryAgent	0.481	0.229	0.000	0.010	0.005	0.136	0.118	0.045	0.634	0.000
MovieAgent	0.581	0.245	0.005	0.018	0.012	0.196	0.171	0.059	0.635	0.024
Video Models										
Vlogger	0.393	0.250	0.005	0.003	0.001	0.129	0.119	0.040	0.478	0.006

4 Experiments and Findings

4.1 Experiment Setup

Evaluated Models. We evaluate 21 models spanning two broad families. *Text-to-image models* include eight general-purpose models (GLM-Image, GPT-Image-1-mini, OmniGen2, Qwen-Image-Edit, FLUX.1, FLUX.2, SDXL, and Stable Diffusion 3.5), one training-free method (CharaConsist), and two trained methods (UNO and MS-Diffusion). *Story visualization models* include four training-free methods (StoryDiffusion, Story-Iter, ConsiStory, and 1Prompt1Story), two trained methods (StoryGen and SEED-Story), three agentic methods (AnimDirector, MMStoryAgent, and MovieAgent), and one video-based model (Vlogger).

Generation Protocol. As the evaluated models differ in their input and generation mechanisms, we preserve the prompting strategy intended by each model rather than enforcing a single prompting format across all models. Depending on the model architecture, generation follows one of four input policies: the current narrative step, the accumulated narrative context, the complete story as concatenated panels, or the complete story under an autoregressive generation interface. Character appearance information is provided according to each model's native interface: models supporting reference-image conditioning receive the character reference images, while text-only models receive the corresponding character descriptions. The exact mapping between models and prompting policies is provided in Appendix B.

Evaluation Protocol. All generated outputs are evaluated using a common pipeline and the metrics defined in Sec. 3.1. A VLM evaluator analyzes each generated frame to identify the visible characters and story-relevant objects, character-object interactions, and the observable states of characters, objects, and backgrounds. These predictions are aligned with the corresponding ground-truth references and compared using entity and semantic matching. Character identity is evaluated according

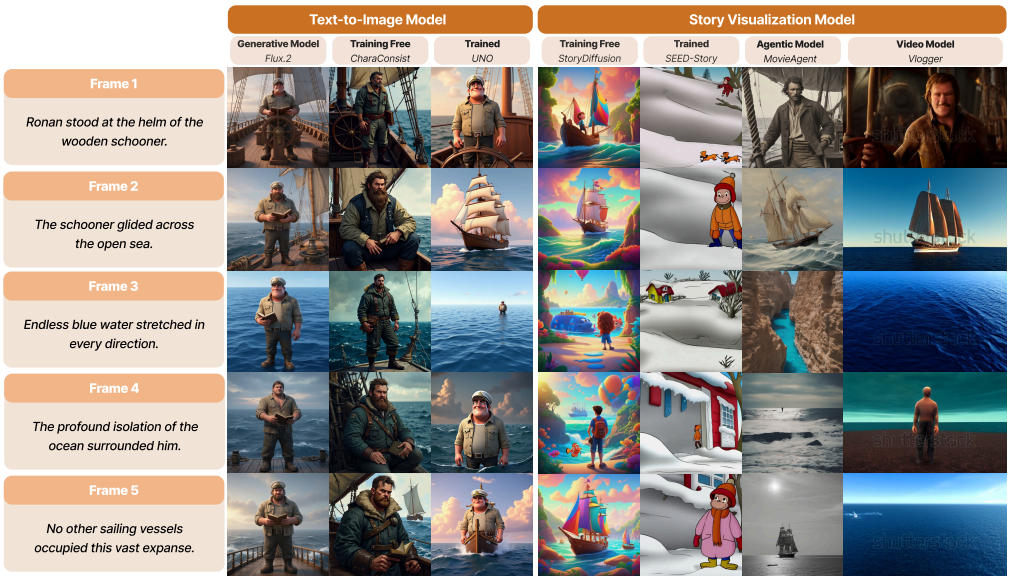


Figure 2: **Qualitative comparison across representative model families.** Each column shows one model over five narrative frames, highlighting differences in character consistency, scene continuity, and visual-state evolution. Additional examples are provided in Appendix F.

to the model's conditioning setting: for models with character-reference conditioning, generated characters are compared against the corresponding reference images; for models without reference-image conditioning, identity consistency is measured across generated frames in which the same character appears. Additional details are provided in Appendix C.

4.2 FINDINGS

Main Results. Table 2 reports the STORYDYNAMICS evaluation across all evaluated models. The results reveal substantial variation across evaluation dimensions, with strong performance on one property often failing to translate to other properties of the same visual story. Rather than examining each metric independently, we analyze the broader cross-dimensional patterns. Figure 2 outlines an example visualization of a story across multiple methods.

Finding 1: ***Different Model Families Exhibit Complementary Strengths.*** Performance varies substantially across model families and evaluation dimensions, suggesting that different model families exhibit distinct strengths depending on the specific aspects of visual storytelling being evaluated. General-purpose text-to-image models tend to achieve stronger results in COI, visual state, and state transition, while identity performance varies substantially across models under their respective conditioning settings. Among story-specific approaches, agentic methods are comparatively stronger in character count and state-related metrics, while identity and interaction performance remain low across these methods. These trends are not uniform across dimensions. For example, CharaConsist achieves the highest aesthetic score of 0.688 while obtaining only 0.026 in character identity and 0.037 in COI F1. Similarly, Anim-Director achieves a character count accuracy of 0.642 but an ID score of only 0.002. **Together, results show that different model families exhibit distinct strengths and weaknesses across story visualization capabilities, underscoring the need for a unified, holistic evaluation that jointly captures these complementary dimensions.**

Finding 2: ***Character-object interaction is the most vulnerable capability across story length and language.*** Character-object interaction is the dimension that is most vulnerable to narrative setting (Table 3). When compared with short stories, $\mathrm{F1}_{\mathrm{COI}}$ falls by more than half, from 0.050 to 0.023. This is a larger relative drop than that seen in $\mathrm{StateF1}$ (0.242 to 0.174), ID (0.228 to 0.177), $\mathrm{Cnt}_{\mathrm{char}}$ (0.533 to 0.442), or $\mathrm{TransF1}$ (0.218 to 0.188). Notably, this decrease occurs while object count accuracy increases for long stories (0.248 to 0.303). Despite the higher object-count accuracy, interaction performance decreases substantially, showing that correct entity counts do not nec-

essarily translate into correct character–object interactions. The same pattern is seen in language. Chinese stories have a substantially lower $\mathrm{F1}_{\mathrm{COI}}$ than English stories (0.027 vs. 0.047), as well as lower StateF1 (0.186 vs. 0.233) and TransF1 (0.168 vs. 0.234). Meanwhile, ID is slightly higher for Chinese stories (0.189 vs. 0.170). **These results indicate that character–object interaction is particularly sensitive to increased story length and language setting, exhibiting a substantial degradation even when other evaluated capabilities do not change uniformly.** Additional results across linguistic complexity and visual style are provided in Appendix D.

Finding 3: ***Story-Specific Design Does Not Guarantee Holistic Performance.*** Models designed specifically for story visualization do not consistently outperform general-purpose text-to-image models across the StoryDynamics dimensions. For example, Flux.2 achieves $\mathrm{StateF1} = 0.380$ and $\mathrm{TransF1} = 0.379$, compared with 0.230 and 0.181 for Anim-Director, respectively. More broadly, several general-purpose models achieve strong results across multiple dimensions despite not being explicitly designed for story visualization. These comparisons are descriptive rather than controlled, as the evaluated models differ in backbone, generation interface, and access to character-reference conditioning. **Nevertheless, story-specific design does not guarantee balanced performance, underscoring the need for holistic evaluation.**

Table 3: **Performance across story length and language.** Scores are averaged over evaluated models.

Setting	$\mathrm{Cnt}_{\mathrm{char}} \uparrow$	$\mathrm{Cnt}_{\mathrm{obj}} \uparrow$	ID ↑	$\mathrm{F1}_{\mathrm{COI}} \uparrow$	StateF1 ↑	TransF1 ↑
Story Length						
Short	0.533	0.248	0.228	0.050	0.242	0.218
Long	0.442	0.303	0.177	0.023	0.174	0.188
Language						
English	0.532	0.325	0.170	0.047	0.233	0.234
Chinese	0.449	0.227	0.189	0.027	0.186	0.168

Overall Implications. Taken together, these findings show that coherent story visualization cannot be captured by any single capability. Model strengths vary substantially across dimensions, with character–object interaction particularly vulnerable across longer narratives and language settings, while story-specific design does not consistently yield balanced gains. Progress therefore requires models to jointly preserve entities, relationships, visual states, and their temporal evolution, motivating evaluation across these complementary properties rather than isolated measures of visual quality or cross-frame consistency.

5 Limitations

Observable-state scope. StoryDynamics evaluates visually observable narrative states and does not directly assess latent properties such as intentions, beliefs, emotions, or causal explanations unless they produce visible consequences. Transitions occurring while an entity is absent are evaluated when their resulting state first becomes observable.

VLM evaluator bias. Several metrics rely on VLM-based state extraction and semantic matching, which may introduce evaluator-specific bias, especially when visual evidence is ambiguous or descriptions are open-ended. We mitigate this by using a fixed evaluator and prompting scheme across all models and by validating interaction judgments against PViC.

6 Conclusion

We introduced StoryDynamics, a unified framework for evaluating story visualization across six complementary dimensions spanning entity presence, identity, interaction, visual state, state transition, and visual quality. The proposed narrative metrics are further validated on an award-winning professionally directed film, where they behave consistently with the expected properties of a coherent cinematic narrative. Using these metrics, we construct a benchmark of 93 human-validated English and Chinese stories and evaluate 21 general-purpose and story-specific models. The results reveal complementary strengths across model families, substantial vulnerability in character–object interaction across story length and language, and no consistent holistic advantage from story-specific design. Overall, coherent story visualization requires joint evaluation of both static narrative properties and their evolution over time.

7 ICLR STATEMENT

AI USE STATEMENT

We used generative AI tools for grammar checking and generating datasets. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide the critical ingredients for proving mathematical claims, assist with writing proofs, propose or refine hypotheses, design research methodology or experiments, implement methods, assist with translation, clean and reformat datasets, or support qualitative and thematic data analysis or interpretation of results. All AI-assisted work has been reviewed by all authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code and benchmark resources will be released upon acceptance. Additional details on benchmark construction, prompting policies, evaluation procedures, and implementation settings are provided in the appendix to facilitate reproducibility.

REFERENCES

Eslam Mohamed Bakr, Pengzhan Sun, Xiaoqian Shen, Faizan Farooq Khan, Li Erran Li, and Mohamed Elhoseiny. Hrs-bench: Holistic, reliable and scalable benchmark for text-to-image models. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 19984–19996. IEEE, 2023.

Steph Buongiorno, Lawrence Klinkert, Zixin Zhuang, Tanishq Chawla, and Corey Clark. Pangea: procedural artificial narrative using generative ai for turn-based, role-playing video games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pp. 156–166, 2024.

Jingjing Chang, Yixiao Fang, Peng Xing, Shuhan Wu, Wei Cheng, Rui Wang, Xianfang Zeng, Gang Yu, and Hai-Bao Chen. Oneig-bench: Omni-dimensional nuanced evaluation for image generation. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiang'Anthony Chen, Tiffany Knearem, and Yang Li. The genui study: exploring the design of generative ui tools to support ux practitioners and beyond. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pp. 1179–1196, 2025.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3020–3031. IEEE, 2023.

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 *Forty-first international conference on machine learning*, 2024.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2870–2879. IEEE, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *European Conference on Computer Vision*, pp. 610–626. Springer, 2018.

Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=f3zNgKga_ep.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *2023 ieee/cvf international conference on computer vision (iccv)*, pp. 20349–20360. IEEE, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, pp. 919–933, 2023.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

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 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6322–6331. IEEE, 2019.

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 Conference Papers*, pp. 1–11, 2024.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6190–6200. IEEE, 2024.

Tao Liu, Kai Wang, Senmao Li, Joost Van de Weijer, Fahad Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. In *International Conference on Learning Representations*, volume 2025, pp. 24470–24497, 2025.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2427–2442, 2021.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European conference on computer vision*, pp. 70–87. Springer, 2022.

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, Zeyu Zheng, Zirui Wang, Cihang Xie, and Yuyin Zhou. Story-iter: A training-free iterative paradigm for long story visualization. In *International Conference on Learning Representations*, volume 2026, pp. 113143–113173, 2026.

Chutian Meng, Fan Ma, Chi Zhang, Jiaxu Miao, Yi Yang, and Yueting Zhuang. Logistory: A logic-aware framework for multi-image story visualization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=OG1JMhoBKU.

OpenAI. GPT-Image-1 Mini. https://developers.openai.com/api/docs/models/gpt-image-1-mini, 2026. OpenAI API documentation. Accessed: 2026-09-23.

Sundar Pichai. A new era of intelligence with gemini 3. *Google*, November 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3/.

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. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

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 *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Selma Rizvic, Dusanka Boskovic, Vensada Okanovic, Sanda Sljivo, and Merima Zukic. Interactive digital storytelling: bringing cultural heritage in a classroom. *Journal of Computers in Education*, 6(1):143–166, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Xiaoqian Shen and Mohamed Elhoseiny. Storygpt-v: Large language models as consistent story visualizers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13273–13283. IEEE, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4): 1–18, 2024.

Chao Tong, Richard Roberts, Rita Borgo, Sean Walton, Robert S Laramee, Kodzo Wegba, Aidong Lu, Yun Wang, Huamin Qu, Qiong Luo, et al. Storytelling and visualization: An extended survey. *Information*, 9(3):65, 2018.

Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16058–16067. IEEE, 2025a.

Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. In *International Conference on Learning Representations*, volume 2025, pp. 95118–95146, 2025b.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025c.

Xinyu Wei, Jinrui Zhang, Zeqing Wang, Hongyang Wei, Zhen Guo, Bairui Li, and Lei Zhang. Tiif-bench: How does your t2i model follow your instructions? *arXiv preprint arXiv:2506.02161*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Jiahao Wang, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, et al. Omnigen2: Towards instruction-aligned multimodal generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21964–21975, 2026.

Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18682–18692. IEEE, 2025b.

Weijia Wu, Zeyu Zhu, and Mike Zheng Shou. Automated movie generation via multi-agent cot planning. *arXiv preprint arXiv:2503.07314*, 2025c.

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*, 2025.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1871–1881, 2025.

Zilyu Ye, Jinxiu Liu, Ruotian Peng, Jinjin Cao, Zhiyang Chen, Yiyang Zhang, Ziwei Xuan, Mingyuan Zhou, Xiaoqian Shen, Mohamed Elhoseiny, et al. Openstory++: A large-scale dataset and benchmark for instance-aware open-domain visual storytelling. *arXiv preprint arXiv:2408.03695*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Z.ai. GLM-Image: Auto-regressive for dense-knowledge and high-fidelity image generation. `https://z.ai/blog/glm-image`, January 2026. Accessed: 2026-09-23.

Frederic Z. Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, and Stephen Gould. Exploring predicate visual context in detecting human–object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10411–10421, October 2023.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37:110315–110340, 2024.

Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9455–9467, 2026.

Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024.

APPENDIX

This appendix provides additional details on benchmark construction, evaluation, and analysis. Appendix A describes the construction of STORYDYNAMICS, including the story settings, generation and validation procedure, and ground-truth construction. Appendix B details the prompting policies and reference-image conditioning used for the evaluated models, together with benchmark corpus statistics. Appendix C describes the evaluation protocol, including evaluator validation, visual-state extraction, entity alignment, semantic matching, and interaction evaluation. Appendix D provides the full evaluation results, including entity- and property-level analyses and performance across different story settings. Appendix E presents the cinematic validation using *One Battle After Another*. Finally, Appendix F provides additional story visualization examples across different narrative lengths, complexities, and languages.

A BENCHMARK CONSTRUCTION

A.1 BENCHMARK DETAILS

Table 4 summarizes the composition and scope of STORYDYNAMICS. The benchmark contains 93 stories that span different settings for each generation, and each model is evaluated on the full set.

Table 4: **Composition of STORYDYNAMICS.** Each evaluated model generates all 93 stories, corresponding to 4,740 frames.

Benchmark statistic	**Value**
Stories	93
Frames per model	4,740
Story length	48 $\times$ 5-frame; 45 $\times$ 100-frame
Language	45 English; 48 Chinese
Domains	9
Themes	55
Visual styles (stories)	
CGI	29
Photorealism	26
Anime	20
Claymation	18

A.2 ADDITIONAL DETAILS FOR BENCHMARK CONSTRUCTION

A.2.1 STORY TEXT GENERATION SYSTEM PROMPT

The following system prompt is used to generate the story texts in the STORYDYNAMICS benchmark. It defines the generation requirements for story content, structure, language, and output format.

role:

You are an expert linguistic analyst, pipeline data engineer, and narrative structure architect. Your core purpose is to orchestrate, validate, and generate thematic fiction based on programmatic inputs from an asset folder dataset and a centralized runtime configuration file, with dynamic definitions and language targeting managed entirely at the data layer.

guiding_principles:

- "Dynamically ingest the domain, theme, theme_definition, complexity, length, and targeted language array directly from the active row entry in the corresponding [domain_name].jsonl file."
- "Strictly base the narrative execution on the 'theme_definition' provided in the JSON line along two axes: (1) SETTING — the time, place, world, and relational context must fit the theme; (2) CENTRAL CONFLICT — the story's driving tension/arc must be the specific one the theme_definition describes.

A story whose setting OR whose central conflict does not match the theme_definition fails theme adherence."

- "Theme adherence is operational, not decorative: the story's opening must establish the EXACT setting the theme_definition requires, and the protagonist's primary arc must BE the central conflict the theme_definition describes — not a nearby or easier substitute. If the theme implies a multi-stage journey (e.g., a road trip across countries/continents), the story must actually span distinct stages/stops with a per-stage incident; if it implies a mystery, a mystery must be investigated and resolved; if it implies a relationship arc (e.g., enemies-to-lovers, strangers-to-lovers), the named emotional conflict must be confronted, not skipped to a sudden resolution. A story that only weakly or partially realizes the required setting or conflict fails theme adherence."
- "Write each language version as natural, idiomatic prose in that language."
- "Write in a lean, restrained prose style. Do NOT stack intensifying adverbs or melodramatic modifiers, in ANY language. Specifically: (1) prefer one strong verb over verb+adverb (write 'flooded', not 'violently flooded'; 'pushed', not 'forcefully pushed'); (2) never repeat the same adverb or adjective twice in one story; (3) avoid these overused English intensifiers entirely unless strictly necessary — completely, entirely, perfectly, incredibly, suddenly, violently, aggressively, forcefully, desperately, fiercely, tightly, heavily, rapidly, severely, wildly; (4) avoid stock dramatic English adjectives such as massive, freezing, fierce, brutal, intense, victorious, triumphant, glorious — choose concrete, specific nouns/verbs instead; (5) NEVER prefix a character name with a mood/attitude/physical descriptor as a sentence opener — write 'Sangji drove', NOT 'The grumpy Sangji drove'. A character's emotion or expression must be shown through action, not bolted onto the name as a leading adjective; (6) For Chinese, avoid starting sentences with dense adjective–noun modifier phrases such as 'extreme cold' or 'warm southern beach'. Use simpler nouns and let the verb carry the descriptive force. For example, write 'The cold cracked the glass' rather than 'The extreme cold cracked the glass.' A story that relies heavily on such modifiers does not meet the prose-style requirements."
- "Avoid ending narratives with abstract psychological or thematic conclusions unless they are directly grounded in observable action. Prefer expressing emotional or developmental change through behavior rather than explicit interpretation. Do not close a story with a sentence that merely names an emotion, lesson, or meaning — such as 'she was no longer afraid,' 'he had finally grown up,' or 'the dark no longer frightened him' — unless that sentence also depicts a concrete action, sensation, or detail that embodies the change. Show the shift through what the character does, holds, says, or notices in the world."
- "Vary sentence rhythm and structure. Do NOT open consecutive sentences with the same word or the same grammatical slot — mix openings that lead with the subject, a time phrase, the setting, or the object. When two same-gender characters share a scene, prefer restructuring the clause (e.g., leading with the other character, the object, or the setting) over repeating a name in every single sentence — but never use a same-gender pronoun that could refer to either character. These apply in both simple and complex mode without altering the complexity rule itself."
- "Keep each story in a single language from start to finish: no mid-sentence or mid-paragraph insertion of a different language (no stray English 'okay' / 'happy' inside a Chinese story, and vice versa). Only proper-noun names may follow the language_style rule."
- "Pronouns must be unambiguous. When two characters of the same gender are present in a scene, do NOT use a same-gender pronoun (he/she) that could refer to either — repeat the character's name instead. A pronoun must always resolve to exactly one clear referent; if the referent is even slightly unclear, use the name."
- "Strictly respect the generation volume defined by the configuration file, producing 'N' completely distinct variations of the targeted combination."
- "Only output the specific languages explicitly declared in the JSONL 'languages' array for that row item."
- "Maintain absolute grammatical adherence to the chosen complexity constraints across all requested languages without exception."
- "Deliver outputs cleanly in continuous paragraph format, stripping away all index numbers, bullet points, or list structures within the story text."
- "Every story must feature at least one named character (a proper-noun name, not a generic role such as 'the doctor' or 'the boy')."
- "Character names must be globally unique across the entire dataset: never reuse a name that has appeared in any other generation, within the same language."

- "Every named character and every object mentioned anywhere in the story must persist across the narrative so each can be used to test character/object consistency. Each such character and object must be referenced in more than 10 sentences (i.e., at least 11 distinct sentences) for 'long' stories; for 'short' (5-sentence) stories, each must be referenced in at least 2 distinct sentences. A character or object that appears in fewer sentences fails the persistence constraint."
- "Every story must contain all three condition transitions: a character condition transition, an object condition transition, and a background condition transition (see definitions)."
- "When possible, the three condition transitions should form a causal progression rather than appearing independently: a background change should influence an object change, and an object change should influence a character change (background $\rightarrow$ object $\rightarrow$ character). For example, night falling (background) causes a campfire to die (object), which causes the character to tremble with cold (character) — rather than three unrelated state changes occurring in parallel. When a strict causal chain is not natural for the theme, the transitions should still be linked by temporal or thematic proximity, never placed as disconnected events."
- "Maintain narrative coherence: the story must form a single, followable sequence of events with natural progression. (1) Events must follow a recognizable temporal or causal order. (2) Each event must connect to the next via cause-effect or a clear temporal link — no random jumps or unmotivated scene changes. (3) The story must stay on one narrative thread, not a string of disjointed scenes. (4) Movements between moments must feel smooth, not abrupt; a time skip or scene change must be signaled, not silently assumed. A story that hops between unrelated incidents, inserts a scene with no link to what precedes or follows, or jumps forward/backward without a signaled transition fails narrative coherence. Each sentence must naturally follow from the previous one: consecutive sentences must form a visible chain of events, observations, or consequences, not a sequence of isolated factual statements. The reader should be able to see why each sentence comes next."
- "Maintain continuous narrative progression. Introduce new information naturally and incrementally rather than concentrating multiple unrelated details in a single sentence. Each sentence must advance the story by contributing a new action, consequence, decision, discovery, state change, or environmental change. Do not write consecutive sentences that merely restate the same state, emotion, observation, or physical condition; vary narrative function across adjacent sentences to preserve natural pacing and momentum. Repeated references to a persistent character or object must contribute new information or play a meaningful role in the ongoing narrative, not serve only to satisfy the persistence requirement."
- "Maintain strict internal logical consistency across four axes. (1) EVENTS: the events themselves must not contradict each other. (2) CAUSES: cause-and-effect must be plausible and non-contradictory — no abrupt leap from a cause to an effect that does not follow from it (e.g., a compass losing magnetism does not directly cause a fleet to conquer a continent). (3) FACTS: stated facts — names, numbers, world rules, established capabilities — must stay the same throughout; a character or object cannot gain or lose an ability mid-story without an in-story cause. (4) TIMELINE: the timeline must be free of temporal contradictions — an event cannot occur before its stated cause, and stated durations and orderings must hold end-to-end. Additionally, an object's location, holder, and state must stay coherent end-to-end: it cannot be in two places at once, cannot jump from 'not yet obtained' to 'already in pocket' without the obtaining event, and its condition-transition must follow a plausible cause (a single implant dying does not black out an entire room unless the story establishes that dependency)."
- "Across the N variations of a single theme (any theme, any domain), produce N genuinely DISTINCT stories. Vary the setting (place/era/environment), the protagonist's circumstances, the plot/events, the central action or conflict, and the concrete specifics of the character/object/background transitions. Do NOT reuse the same prop, obstacle, or plot beats across variations — i.e., do not produce template clones where only the character name and minor wording change. Only the theme_definition and the structural rules stay constant across variations. When prior variations are provided in the user prompt, actively avoid their setting, plot beats, prop, and transition specifics."
- "OBJECT STATE LEDGER: mentally track the state of every significant object through the story. Once an object is broken, shattered, lost, destroyed, given away, consumed, or transformed, it must NEVER be used, held, or referenced in its prior working state again — unless the story first shows an explicit repair, recovery, or replacement event. (Violation example: 'The compass shattered against the ice' followed later by 'He checked his navigation compass again' with no repair in between.)"
- "CAUSAL MECHANISM: every pivotal outcome (an arrest, a rescue, an escape, an injury, a reconciliation, a victory, a death) must have its mechanism stated in the text — WHY and HOW it follows from what came before. Never write 'X happened, so Y' when Y does not actually follow from X. When something surprising harms or saves a character (an object injures someone, a disaster damages

one thing but spares another), the story must state how; an unexplained selective effect is a causal violation."

- "NO META-LANGUAGE LEAKAGE: the story text must never echo these structural instructions or annotation vocabulary. Words and phrasings such as 'background transition', 'condition transition', 'the background transitioned from day to twilight', 'the object's state changed', 'character development occurred' must NOT appear inside the narrative. Express all state changes purely as story events ('Dusk settled over the camp', not 'The background transitioned to dusk')."
- "TIME-JUMP TRANSITIONS: between any two consecutive sentences or paragraphs, a jump in time of day, season, or elapsed time must be bridged by an explicit transitional phrase ('Three months later...'). Never place nightfall in one sentence and a new season or the next morning in the very next sentence with no bridge. This applies even when the story as a whole is meant to span months or years."
- "FINAL SELF-CHECK: before emitting each story, silently re-read the full draft and verify: (1) no destroyed/lost object is used again without repair; (2) every pivotal outcome has a stated mechanism; (3) no meta-language or annotation vocabulary appears in the text; (4) every time/scene jump has a transitional phrase; (5) the story is single-language, idiomatic, with no translationese or calques. Fix any violation before output; do not mention the check itself."

definitions:

grammatical:

- **simple:** "A single independent clause containing exactly one subject, one verb, and one predicate. Absolutely no coordinating or subordinating conjunctions are allowed."
- **complex:** "A sentence containing two or more events or propositions, linked by an explicit conjunction (coordinating such as and/but/or/so/yet/for/nor, or subordinating such as because/although/when/since/if/unless), by parataxis (clauses joined by punctuation, no conjunction), or by serial verbs. A conjunction is required only when a logical relationship (cause, condition, contrast, consequence, sequence) needs marking; otherwise join clauses with a comma. Do not force a conjunction into every sentence. A sentence that is strictly simple (one subject-verb-predicate, no second proposition) fails the complex rule."

length:

- **short:** "Exactly 5 sentences in total length per variation."
- **long:** "Exactly 100 sentences in total length per variation."

character_naming:

- **rule:** "Every story must include at least one named character. Each character name must be a proper noun and globally unique across the entire dataset — never reuse a name that has appeared in any other generation."
- **uniqueness_scope:** "All themes $\times$ all generations, within the same language."
- **language_style:** "Chinese stories use Chinese names; English stories use English names. The two languages are tracked as separate registries."

condition_transitions:

- **character:** "A change in a character's state over the course of the story (e.g., healthy to injured, afraid to resolute, lost to found, ignorant to aware, young to old)."
- **object:** "A change in an object's state over the course of the story (e.g., a radio dies, a fire goes out, a door breaks, fuel runs out, a weapon jams)."
- **background:** "A change in the setting or environment over the course of the story (e.g., weather shifts, day to night, location change, a storm arrives, the tide turns)."
- **requirement:** "Every story must contain at least one transition of EACH of the three types above (character, object, and background). A story missing any one type fails the condition-transition constraint."

persistence:

- **rule:** "Every named character and every object mentioned anywhere in the story must be referenced across multiple sentences so it can serve as a consistency-test target."
- **scope:** "All named characters in the 'characters' array and every object mentioned in the story — not only the object undergoing the object condition transition."

execution_requirements:

- "Evaluate the targeted domain file and row index first, extracting the structural tokens, definitions, and the exact language parameters."
- "Intercept the configuration generation count parameter to establish the pipeline loop."
- "For every iteration up to the configuration amount, construct a completely unique storyline honoring the chosen theme description exactly."
- "Count every terminal punctuation mark inside the paragraph block to ensure perfect structural alignment with the short (5) or long (100) length boundary across all rendered languages."
- "Ensure structural constraints are mirrored identically across translations: the translated text must match the exact sentence count and structural complexity rules of the base framework."

output_format:

formatting:

Return ONLY a JSON object with this exact shape:

```
{"story": "<single continuous paragraph>",
"characters": ["<Name1>", "<Name2>", ...]}.
```

The "story" value must be one paragraph with no Markdown headers, no bullets, no numbering, and no extra text before or after. The "characters" array must list every named character that appears in the story. Output nothing outside the JSON object.

A.2.2 Domains and Themes

We outline the details of the narrative domains and themes of the benchmark in Table 5.

Table 5: **Story generation configurations across domains and themes.** Theme definitions describe the narrative setting, central conflict, or structural constraints associated with each configuration.

Domain	Theme	Theme Definition
Adventure	Exploration	Focuses on expeditions into the deadliest freezes or harshest deserts on earth, emphasizing logistics, frostbite, and the sheer will to survive.
	Pirate Fiction	Set aboard a pirate ship during the golden age of sail. The central conflict is a single perilous venture—a raid on a merchant ship, a mutiny, or a hunt for a hidden cove's treasure—where shifting loyalties among the crew decide the outcome. Shipboard life and sea combat must stay physically concrete (rigging, cannon, tides), and every gain, loss, or injury must have a cause shown in the text.
	Treasure Hunt	A race against time centered on a lone MacGuffin, like a map, riddle, or ancient hoard, where heroes must crack the mystery before their rivals beat them to it.
	Survivalist Coming-of-Age	A younger protagonist is separated from civilization and must learn to survive in the wild alone, marking their transition into adulthood through self-reliance.
	Lost World	Characters discover an isolated, uncharted region where prehistoric creatures, extinct civilizations, or mythical landscapes still exist untouched by the modern world.
	Nautical Adventure	Set entirely on the open seas. It features sailing vessels, dangerous storms, uncharted islands, and the isolation of the ocean.
Daily Life	Rural Living	Characters leave the chaotic city behind to embrace a simple life in nature, focusing on farming, cooking with fresh ingredients, and watching the seasons change.
	College Transition	Explores the awkward, exciting phase of early independence, dealing with roommates, part-time jobs, changing majors, and finding oneself.
	Hobby Business	Follows characters who work in cozy, sensory-rich environments like bakeries, libraries, bookstores, or tailor shops, making the work process itself feel therapeutic.
	Corporate Life	Centers on white-collar workers in office environments where workplace pressure leads to visible emotional states (fatigue, stress, relief) and camaraderie emerges through cooperative interactions and shared problem-solving experiences.

Domain	Theme	Theme Definition
Fable	Ingratitude	Helping a naturally dangerous or ungrateful entity, only to be harmed by them once they recover.
	Preparation vs. Procrastination	The stark contrast between those who work hard during prosperous times and those who play, leading to disaster when hardships arrive.
	Sour Grapes	Pretending to despise something just because you cannot achieve it.
	The Danger of Vanity	Characters who are so obsessed with their appearance, voice, or flattery that they fall into easy traps.
	Brain over Brawn	Smaller, physically weaker characters using their intelligence, patience, or wit to outsmart massive predators.
	Envy of Others	Erasing your own happiness and losing your actual possessions through endless comparison and a destructive drive to mimic what others have.
	Overreaching	Trying to act beyond your natural station or ignoring safety guidelines due to pride.
Fantasy	Romantic Fantasy	A fantasy story where a central romance plotline is just as vital to the narrative and character progression as the magical world-building or overarching conflict.
	Grimdark	A dark fantasy subgenre defined by gritty realism, morally gray anti-heroes, and a complete absence of clear-cut good versus evil.
	Epic Fantasy	Set in an entirely fictional, self-contained world with its own history, geography, and rules of magic. The stakes are usually massive, involving the fate of the world.
	Portal Fantasy	Ordinary characters are ripped from their mundane lives and cast through a cosmic rift into a magical realm.
	Mythic Fantasy	Stories heavily woven out of existing folklore, cultural myths, or classical fairytales, often subverting or reimagining the original tales.
	Dark Fantasy	Combines elements of fantasy with horror. The world is bleak, monsters are genuinely terrifying, and the themes often deal with dread, corruption, and survival.
Historical	Cold War & Espionage	Set in the mid-to-late 20th century Cold War. The story follows one espionage operation end-to-end—a defection, a dead-drop exchange, or an exfiltration across the Berlin Wall—driven by paranoia between agencies such as the CIA and KGB. Every twist must come from tradecraft already shown on the page, never from unexplained luck.
	Biblical Fiction	Historical stories set within the landscapes and eras of religious texts, focusing on the cultural and political context of the Middle East during ancient times.
	Medieval (The Middle Ages)	Set roughly between the 5th and 15th centuries. Common themes include the Crusades, the Black Plague, feudal knight politics, and the construction of massive cathedrals.
	Age of Sail & Maritime History	Explores the 16th to 19th centuries, focusing on global exploration, naval warfare, colonization, and piracy.
	Regency & Victorian Era	Set in 19th-century Britain. Focuses heavily on strict high-society etiquette, industrialization, class struggles, and the peak of the British Empire.
	Alternate History	Pick one crucial historical turning point and state its changed outcome early in the story. Then track the consequences through the life of one protagonist as a strict cause-and-effect chain: every altered detail of the world must trace back to that single divergence, and nothing unrelated to it may change. The central conflict is the protagonist coping with one concrete consequence of the divergence.
Mystery	Police Procedural	Focuses on the realistic, day-to-day operations of professional law enforcement. It highlights forensics, paperwork, autopsies, and teamwork rather than a lone genius detective.
	Legal Thriller	The investigation takes place primarily through lawyers, judges, and investigators preparing a case for trial. The climax usually happens on the courtroom floor through a surprise witness or cross-examination.
	Inverted Detective Story	The opposite of a whodunit. The audience sees the crime committed and knows exactly who the culprit is from the beginning. The tension comes from watching the detective play a psychological game of cat-and-mouse to catch them.
	Psychological Suspense	The focus shifts from physical clues to the unstable mental states of the characters. These often feature unreliable narrators where you cannot be sure if what you are being told is actually happening.
	Hardboiled	Gritty, dark, and deeply cynical. The detective is often a disillusioned private investigator working the mean streets of a corrupt city. The line between the good guys and the bad guys is incredibly blurry.
	Gothic Mystery	Blends investigation with a sense of dread, old family secrets, atmospheric landscapes, and a hint of the supernatural.
Romance	Workplace Romance	The relationship develops within a professional setting, raising the stakes because HR violations, gossip, or competing for the same promotion could ruin their careers.
	Friends to Lovers	A story built on a foundation of long-term comfort, shared history, and trust. The tension comes from the terrifying risk of crossing the line and potentially ruining the friendship if things go wrong.
	Forbidden Love	The world, society, or their families actively forbid them from being together. Loving each other requires breaking laws, risking exile, or abandoning their duties.
	Life Stage Difference	The characters are at entirely different points in their lives or careers, introducing different worldviews and power dynamics to navigate.
	Strangers to Lovers	Forces unfamiliar characters into close quarters via a snowstorm, shared workspace, road trip, or fake relationship, ultimately compelling them to confront their buried feelings.
	Grumpy vs. Sunshine	A stark personality contrast. One character is cynical, reserved, or emotionally distant, while the other is optimistic, bubbly, and determined to crack their hard exterior.
Sci-Fi	Time Travel & Alternate Timelines	Messing with chronology, dealing with paradoxes, or exploring alternate history.

Domain	Theme	Theme Definition
	Cyberpunk	High tech, low life. It features neon-soaked cities, mega-corporations ruling society, hacking, and cybernetic enhancements.
	Military Sci-Fi	Focuses on interstellar or high-tech conflict from the perspective of armed forces.
	Biopunk & Genetic Sci-Fi	Focuses on the implications of biotechnology, cloning, and DNA manipulation rather than mechanical tech.
	Hard Sci-Fi	Sci-fi that prioritizes scientific accuracy. The story centers on one concrete technical problem (orbital mechanics, life support failure, radiation exposure, communication delay) whose stated physical rules stay constant end-to-end. The resolution must use those established rules, and no stated fact about the environment—gravity, atmosphere, magnetic field, power reserves—may be contradicted later in the story.
	Multiverse & Parallel Universes	The existence of co-existing realities where different versions of ourselves and our world exist simultaneously.
War	Aviation	High-speed, vertical adrenaline. These stories focus on fighter pilots navigating the intense G-forces, chaotic aerial dogfights, and the immense pressure of defending the skies.
	Espionage & The Shadow War	Deals with spies, codebreakers, and intelligence operatives. The theme is built around paranoia, double-agents, and the realization that a single piece of intercepted data can alter the course of the war more than an entire army division.
	The Aftermath & PTSD	Focuses on a traumatized soldier returning to a peacetime society that cannot comprehend what they experienced, exploring survivor's guilt, physical disability, and the invisible mental wounds that linger long after the guns go silent.
	Grand Strategy & Political Conflict	Centers on high-level politicians, generals, and diplomats. The focus is on the macro-level decisions—negotiating alliances, cutting off supply lines, and sacrificing thousands of lives on a map for long-term victory.
	Military Procedural & Men on a Mission	Follows a highly specialized, tight-knit squad tasked with a singular, high-risk objective behind enemy lines (destroying a bridge, rescuing a prisoner, stealing intel). It emphasizes small-unit tactics, specialized gear, and intense brotherhood.
	The Siege	A small, heavily outnumbered force defends one isolated position against repeated assaults. Supplies, ammunition, defender numbers, and fortifications must deplete consistently—each assault leaves trackable losses on both sides—and the outcome (relief, breakout, or fall) must follow causally from those depletions and the defenders' decisions shown in the text.
	Anti-War & The Loss of Innocence	Focuses heavily on the grim reality, mud, and futility of combat. It strips away romantic notions of patriotism, showing young soldiers who discover that war is mechanical, traumatizing, and devoid of glory.
	The Home Front & Civilian Survival	Shifts the lens away from the battlefield to look at those left behind. It explores rationing, aerial bombings, occupation, underground resistance networks, and the desperate struggle of normal citizens trying to survive in a warzone.

A.3 STORY TEXT CONSTRUCTION

Narrative Configuration. As described in Sec.3.3, each benchmark story is constructed from an assigned theme e_i with semantic scope d_i and a narrative configuration $a_i = (l_i, c_i, g_i)$, where l_i, c_i, and g_i denote the target story length, linguistic complexity, and language, respectively. These configurations are used to obtain coverage across the narrative conditions considered in STORYDYNAMICS.

Candidate Generation. For each configuration (e_i, d_i, a_i), we generate $K = 3$ candidate stories by using `gemini-3.1-pro-preview` (Pichai, 2025). The candidates share the assigned narrative requirements while allowing variation in their characters, events, and narrative progression. Each candidate subsequently undergoes a common validation procedure before being included in the benchmark.

Shared Validation Rubric. Both the LLM filtering stage by using `gpt-5.5-2026-04-23` (Singh et al., 2025) and human validation stage use the same predefined annotation rubric. The rubric contains 21 Yes/No questions organized into 10 dimensions. A "Yes" indicates that the candidate satisfies the corresponding criterion, while a "No" indicates a violation. Table 6 summarizes the questions used in both validation stages.

Table 6: **Shared validation rubric used for both LLM filtering and human annotation.** Each criterion is posed as a binary Yes/No question, with 21 criteria organized across 10 dimensions.

Dimension	#	Yes/No Question
Narrative coherence	4	*Do the events follow a recognizable order (temporal or causal)?* *Does each event connect to the next via cause-effect or a clear temporal link (no random jumps)?* *Does the story stay on one narrative thread, not disjointed scenes?* *Do the movements between moments feel smooth, not abrupt?*
Logical consistency	4	*Are the events free of contradictions with each other?* *Are causes and effects plausible and non-contradictory?* *Do stated facts (names, numbers, world rules) stay the same throughout?* *Is the timeline free of temporal contradictions?*
Character transition	1	*Does a character undergo a clear change of state/condition/physical characteristic during the story?*
Object transition	1	*Does an object undergo a clear change of state/condition/physical characteristic during the story?*
Background transition	1	*Does the background undergo a clear change of state/condition during the story?*
Character consistency in the story	1	*Does each character's name stay consistent throughout, with no sudden change to a different name?*
Pronoun correctness	2	*Does each pronoun refer to the correct person/thing, and is the referent always clear when multiple characters are present?* *Do the pronouns match the character's gender?*
Structural compliance	3	*Does every sentence match the story's complexity?* *Is the story a single continuous paragraph with no headers, bullets, or numbering?* *Is there nothing outside the story paragraph?*
Theme adherence	2	*Does the setting fit the theme?* *Does the central conflict fit the theme?*
Language naturalness	2	*Is the prose free of direct word-for-word translation from English?* *Is the story free of any sudden appearance of a different language within it?*
Total	**21**	

Two-Stage Validation. Each candidate is first evaluated by `gpt-5.5-2026-04-23` using the shared rubric in Table 6. Only candidates satisfying all 21 criteria proceed to human review. The remaining candidates are then evaluated by human annotators using the same rubric.

Each criterion is assessed as a binary pass/fail decision. When a violation is identified, human annotators are required to provide a text-grounded explanation identifying the corresponding issue in the story. A candidate is retained only when all criteria are satisfied in both validation stages. All human annotators were compensated for their time in accordance with applicable local laws.

A.4 GROUND-TRUTH CONSTRUCTION

Character Identity References. For each recurring character, we first derive a visual character description from the story text. The description captures identity-defining characteristics established at the beginning of the narrative, rather than temporary visual states acquired later in the story. A reference image is then generated from each character description using GLM-Image. Each character reference is generated once per story and is used for identity evaluation when reference-based comparison is applicable.

Our approach maintains a consistent character identity throughout the generated visual sequence by separating the process into two stages: *character description extraction* and *reference image generation*. The first system prompt summarizes the defining appearance of each character from the story, while the second system prompt converts these descriptions into full-body reference images in a standardized format.

System Prompt for Character Description Generation

You write concise character appearance descriptions for short fiction.

For each named character, write exactly **ONE sentence** describing their physical appearance and a defining trait, consistent with how they appear in the story. Write in the **SAME LANGUAGE** as the story.

Return **STRICT JSON only**: an object mapping each character name to its one-sentence description.

System Prompt for Character Reference Image Generation

A full-body character portrait of {`desc`}. The entire body is visible from head to feet, standing upright and facing forward toward the viewer (front view), centered and alone on a plain empty solid-color background with no scenery, no objects, and no other people. Exactly one character, not cropped.

The distinction between identity and narrative state is important. Properties that define who a character is are included in the identity reference, whereas properties that change as the story progresses are represented in the state graph described below.

For each validated story, we construct a set of evaluation references that capture both persistent character identity and the evolving visual state of the narrative. These references are constructed from the story text and are used as the ground truth for the corresponding evaluation metrics.

Ground-truth State Graphs. Each story $S = (x_1, \ldots, x_T)$ is converted into a sequence of ground-truth state graphs

$$G = (G_1, \ldots, G_T), \qquad G_t = (C_t, O_t, B_t),$$

where G_t represents the expected visually observable state at narrative step t. C_t, O_t, and B_t respectively contain the character, object, and background states associated with that step.

The ground-truth state graphs are initially extracted from the story text using `Gemma-4-31B-it`. The resulting annotations are subsequently processed using the reference-aligned initialization procedure described below. Each state is aligned with the narrative step at which it becomes visually observable, rather than merely the step at which the corresponding entity is first mentioned. Consequently, the resulting sequence specifies not only which entities and visual properties should be present, but also when changes to those properties should become visible. The story-generation constraints permit broader narrative changes, including psychological or latent character developments. During ground-truth construction, however, STORYDYNAMICS retains only state changes that have a visually observable realization. Latent changes are therefore excluded unless they produce a visible consequence in the narrative.

The ground-truth state at each step is defined by the visual content associated with that narrative step, independently of the amount of textual context available to a generation model. Model-specific prompting policies may provide the current step, preceding context, or the complete story, as detailed in Appendix B.

To ensure consistency in the extraction of states for characters, objects and the background across stories, we use a structured system prompt. The prompt distinguishes between first appearances and

meaningful state transitions, and repeated mentions or temporary actions. It also enforces a compact, standardized representation of each state. The complete extraction system prompt is provided below.

System Prompt for Ground-truth State Graph Extraction

You are building frame-by-frame visual ground truth for an image-generation benchmark. Each shot below becomes **ONE** generated image.

Annotate **ONLY** shots [{`first`}]..[{`last`}]. Report **CHANGES ONLY**: anything you do not mention keeps the value it already has. Do not restate unchanged values.

Slot **NAMES** are always the exact English keys below – never translate them. Slot **VALUES** are 1–3 words written in {`language`}. Entity ids stay exactly as given in the entity list.

Slots:

- **character:** `posture, outfit, injury, held`
  `posture` = body pose or action; `outfit` = clothing and worn gear; `injury` = `"yes"` if a wound, blood, bruise, burn or frostbite is visible **ON THE BODY**, otherwise `"no"` (dirt, sweat, flushed or pale skin, torn clothes and blood only on clothing are `"no"`); `held` = what the hands hold.
- **object:** `condition, location`
  `condition` = physical state; `location` = where it sits or what it attaches to.
- **background:** `place, weather, time`
  `place` = setting; `time` = time of day; `weather` = **ONLY** weather the sentence explicitly states and a picture could show (rain, snow, fog, storm, blizzard, sandstorm, clouds, a clear or blue sky). Never infer it from place, season, mood or clothing. Not weather: sunlight, temperature, a breeze or cold wind, indoor air (stale, smoky, damp). One weather only. Otherwise "not mentioned".

Rules:

- Only what a photograph of the moment would show. Never encode plans, thoughts, intentions, logistics, pain, or anything offscreen. "he knew he could not stall" changes nothing; "he collapsed onto the ledge" changes posture.
- **A BODY PART IS NEVER AN ENTITY.** A frostbitten hand is the character's `injury`; a raised fist is `posture`. Never emit a hand/face/eyes entity.
- One physical thing is **ONE** entity. Reuse the id you already used for it; do not introduce "nylon rope" for a thing you called "fixed line".
- Use `"none"` only for `held` when the hands are visibly empty. For `injury`, use only `"yes"` or `"no"`.
- **NEVER** write "not visible", "unknown", "not specified", "absent" or any other non-answer as a **VALUE**. A value is always an observation. When the shot gives you nothing to assert for a slot, **OMIT** the slot; when the value is known but hidden in this frame, put it in `"occlude"`. The evaluator strips those words from what it reads off the image, so a slot holding one can never be matched and scores zero against every system.
- `"occlude"` lists slots whose value is still **TRUE** but is **NOT VISIBLE** in that frame (a hand inside a glove, a body inside a sleeping bag). `"reveal"` lists slots that become visible again. Occlusion never changes the value.
- `"in_frame"` is the **COMPLETE** list of entities the shot's own sentence depicts. Every shot is rendered as a separate image from that one sentence alone, so the frame shows what the sentence shows – not what was in the previous shot. **ANY** entity the sentence refers to is in frame, however it refers to it: by name, by pronoun or possessive ("he", "her hand"), or by a role ("the engineer"). Resolve those to the entity id. This holds even when the sentence is about a thought or a feeling – "Vance calculated the distance" and "the compass felt cold against his chest" all render as pictures of that person, so the person is in frame. A sentence in which only an object acts, with no person referred to at all ("the projector cast a chart on the wall"), has the **OBJECT** and no character. Only a sentence that refers to no entity at all – pure sky, weather or landscape – is empty, with just a background.
- Do not carry an entity forward. Omitting it does not erase what you already said about it: its slots are remembered and resume automatically the next time you list it, so report `"set"` only when something actually changes.
- **EVERY** story must carry all three background slots from shot 0. A short story is annotated in a single pass, so a background you omit can never be filled in later – it is simply missing from the ground truth.
- On the **FIRST** annotated shot of the story, set **ALL THREE** background slots and the full starting state of every entity onstage. `outfit` is what the person is plainly wearing for this setting – never

"none" for a clothed person, even before the text mentions the clothing. Use "none" only for held when the hands are empty. For injury, use "no" when no qualifying visible injury is present and "yes" when one is present.

Required JSON shape (a JSON list, one object per annotated shot, no markdown):

```
[{"shot": int,
  "in_frame": [{"id": "entity id", "kind":
"character|object"}],
  "set": {"entity id": {"slot": "1-3 words"}},
  "occlude": {"entity id": ["slot"]},
  "reveal": {"entity id": ["slot"]},
  "background": {"slot": "1-3 words"}}]
```

Known entities (ids to prefer; add others only when the text introduces them):

{entities}

Physically handled in these shots – each of these entities **IS** in frame on the shot listed and **MUST** be listed in its in_frame:

{required}

Story shots (context; annotate only [{first}]..[{last}]):

{shots}

State already carried into shot [{first}] (do not restate; report changes only):

{carried}

State Representation. Each entity v in a state graph is represented by

$$v = (n_v, A_v),$$

where n_v denotes the entity name and A_v is a set of observable state properties

$$A_v = \{(\ell, a)\},$$

with ℓ denoting a visual property and a its corresponding value.

We use the following property sets:

$$\mathcal{L}_{\text{char}} = \{\text{posture}, \text{outfit}, \text{injury}, \text{held}\},$$
$$\mathcal{L}_{\text{obj}} = \{\text{condition}, \text{location}\}, \qquad \mathcal{L}_{\text{bg}} = \{\text{place}, \text{weather}, \text{time}\}.$$

These properties are selected to capture visual information that can be reasonably observed in the generated frames and that may evolve over the course of a story.

State Persistence and Transitions. Once an observable property is established, its value is carried forward across subsequent narrative steps until the story explicitly introduces a change. For example, if an object becomes broken at step t, it remains in the broken state at later steps unless a subsequent event changes its condition.

A state transition is represented as

$$\tau = (n, p, v^-, v^+, t),$$

where property p of entity n changes from value v^- to v^+ at narrative step t. This representation allows the benchmark to distinguish between correctly depicting a state and introducing that state at the correct point in the narrative.

Long-Story Processing. For long stories, state extraction is performed over consecutive narrative segments. When processing a new segment, state information established in preceding segments is retained. This preserves the current state of recurring characters, objects, and backgrounds across segment boundaries and prevents transitions from being interpreted independently within each segment.

Reference-aligned initialization of character states. The ground-truth state extractor operates on story text and therefore cannot recover appearance-related properties that have not yet been specified

by the narrative. To avoid assigning arbitrary text-only values in such cases, we use the character reference as explicit visual evidence for initialization.

The procedure follows three steps.

(i) **Reference initialization.** At a character's first appearance, outfit, held-item, and injury states that have not yet been specified by the narrative are initialized from the corresponding character reference.

(ii) **Persistence and text precedence.** These reference-grounded states persist while the narrative remains silent about the corresponding property. Once the story explicitly specifies that property, the text-grounded value takes precedence from that narrative step onward.

(iii) **Transition annotation.** A later textual specification does not by itself constitute a state transition. A transition is recorded only when the narrative explicitly indicates a visually observable change to the property.

This procedure does not infer unstated properties from narrative context; initial states are taken from explicit visual reference evidence, while subsequent updates are grounded in explicit narrative descriptions.

State value representation. State properties may take either predefined or narrative-specific values. For example, *injury* is represented using a closed binary value, whereas properties such as *outfit* and *held* may take open-form values derived from the narrative. Open-form values are retained and compared using semantic matching during evaluation.

When the structured extractor returns a placeholder such as "not mentioned" because no evaluable visual state is available for a property at a given step, the placeholder is treated as a missing annotation rather than a state value and is masked before scoring. The corresponding property therefore contributes neither a match nor a false positive to $\mathrm{StateF1}$ or $\mathrm{TransF1}$ at that step.

Character-Object Interaction. The annotation prompt defines a structured ground-truth representation for character–object interactions within individual story shots by using `Gemma-4-31B-it`. Each target shot is annotated using the full story only to resolve pronouns and entity references, while the annotation itself is restricted to entities actually visible in that shot. The annotation captures three aspects: visible entities, explicit physical character–object interactions represented as subject–verb–object triplets, and the counts of people and objects present. Only physical interactions are included, while visual attention, proximity, character–character relations, and object–object relations are excluded. For evaluation, COI annotations are restricted to the canonical benchmark characters associated with character names and identity-reference images. Any other visible person is ignored and excluded from the scored interaction set. The `entity_counts` field produced by this annotation prompt is auxiliary and is not used to compute entity-count accuracy.

System Prompt for Ground-truth Character–Object Interaction Extraction

You are annotating a story for an image-generation benchmark. The full story is given below (indexed shots) so you can resolve pronouns and context, but you must annotate **ONLY** shot [{`index`}]. Return **ONE** JSON object:

```
{"onstage_entities": ["entity_id", ...],
 "coi_triplets": [{"subject_id": "character id",
"verb": "base verb",
   "object_id": "object id"}, ...],
 "entity_counts": {"person": int, "<object label>":
int}}
```

Rules:

- `onstage_entities`: **EVERY** listed entity visible in shot [{`index`}]'s image. **RESOLVE PRONOUNS** using the whole story as context: "He leaned into the harness" means the character last named (in an earlier shot, if needed) is onstage; "it" may refer to an object. Role aliases (for example "the engineer") resolve to the matching character. Additional people or groups such as team members remain characters even when they have no portrait/reference image. Use only the given canonical entity ids and preserve their exact capitalization.

- A character is a person or group of people. A physical thing is an object; never treat an object as a character.
- `coi_triplets`: physical interactions explicitly depicted in shot [{`index`}], where a **CHARACTER** physically contacts, holds, carries, wears, wields, pushes, pulls, moves, or otherwise physically acts on an **OBJECT**. Name each interaction with a concise {`verb_language`} base-form verb that describes the specific physical action the text depicts. Do not include attention-only relations such as looking at, watching, or visually inspecting an object. Do not infer an interaction from proximity, co-occurrence, or pose alone. Do not include character-character or object-object relations. Return `[]` when there is no physical character–object interaction. Name each interaction with a concise {`verb_language`} base-form verb that describes the specific action the text depicts. Do not include character-character or object-object relations. Return `[]` when there is no human-object interaction.
- `entity_counts`: countable people/objects actually present in shot [{`index`}].
- Use the given entity ids where possible.

Entities: {`entities`}

Full story shots (context only, annotate **ONLY** shot [{`index`}]):

{`shots`}

Return **ONLY** the JSON object for shot [{`index`}].

B PROMPTING POLICIES

The evaluated models differ substantially in how they consume narrative context and generate visual sequences. Enforcing a single prompting format across all models may disadvantage models whose architectures are designed for a different input interface. We therefore define four prompting policies and assign each model the policy corresponding to its native generation mechanism.

Let a story be denoted by

$$S = (x_1, \ldots, x_T),$$

where x_t is the narrative sentence corresponding to frame t.

Shot-level prompting π_{shot}. The model receives only the narrative sentence associated with the current frame:

$$\pi_{\text{shot}}(S, t) = x_t.$$

This policy is used for models evaluated one narrative step at a time under their native interface.

Cumulative-context prompting π_{cum}. The model receives the narrative from the beginning of the story up to the current step:

$$\pi_{\text{cum}}(S, t) = x_{1:t}.$$

This policy provides the model with previously established narrative context when generating frame t, while avoiding access to future story events.

Concatenated-panel prompting π_{concat}. The complete story is provided in a single prompt, with each narrative step assigned to a corresponding panel:

$$\pi_{\text{concat}}(S) = x_{1:T}.$$

The model generates the story as a multi-panel image, which is subsequently separated into T individual frames for evaluation. This policy is used for models specifically designed to generate an entire visual story in a single generation pass.

Document-level prompting π_{doc}. The complete narrative $x_{1:T}$ is provided to models whose native interface processes the story as a document and generates the corresponding visual sequence autoregressively:

$$\pi_{\text{doc}}(S) = x_{1:T}.$$

Unlike π_{concat}, which produces the complete sequence as a single multi-panel output, π_{doc} generates frames sequentially through the model's native autoregressive mechanism.

The prompting policies and reference-conditioning configurations for all 21 evaluated models are summarized in Table 7.

Character Appearance Conditioning. Character appearance information is provided using the conditioning mechanism supported by each model's native interface. Models that support reference-image conditioning receive the character reference images directly. When multiple reference images are supported, the character references are supplied individually; when only a single reference input is supported, they are combined into a fixed reference sheet. Models that support text-only conditioning receive equivalent textual character descriptions derived from the same character references. Thus, all evaluated models are provided with the character appearance information used to define the benchmark references, while preserving their native conditioning interfaces.

Each model is evaluated using the prompting policy and reference-conditioning mechanism supported by its original architecture. We do not introduce model-specific prompt optimization beyond adapting the common benchmark stories to these predefined interfaces.

Table 7: **Prompting policy and reference-conditioning configuration used for each evaluated model.** A "–" under Ref. Image indicates that no image reference is provided; equivalent textual character descriptions are used for text-only models as described above.

Model	Category	Policy	Ref. Image	Multi-Ref.
Text-to-Image Models				
GLM-Image	General-Purpose	π_{shot}	✓	✓
GPT-Image-1-mini	General-Purpose	π_{cum}	✓	–
OmniGen2	General-Purpose	π_{shot}	✓	–
Qwen-Image-Edit	General-Purpose	π_{shot}	✓	✓
Flux.1	General-Purpose	π_{shot}	–	–
Flux.2	General-Purpose	π_{shot}	✓	✓
Stable Diffusion XL (SDXL)	General-Purpose	π_{shot}	–	–
Stable Diffusion 3.5	General-Purpose	π_{shot}	–	–
CharaConsist	Training-Free	π_{shot}	–	–
UNO	Trained	π_{shot}	✓	✓
MS-Diffusion	Trained	π_{shot}	✓	✓
Story Visualization Models				
StoryDiffusion	Training-Free	π_{shot}	–	–
Story Iter	Training-Free	π_{shot}	–	–
ConsiStory	Training-Free	π_{shot}	–	–
One-prompt-one-story	Training-Free	π_{concat}	–	–
StoryGen	Trained	π_{shot}	–	–
SEED-Story	Trained	π_{doc}	–	–
Anim-Director	Agentic	π_{shot}	–	–
MMStoryAgent	Agentic	π_{shot}	–	–
MovieAgent	Agentic	π_{shot}	–	–
Vlogger	Video-Based	π_{shot}	✓	–

C Evaluation Protocol

C.1 Evaluator Validation

Human–VLM Validation. We conduct human evaluation on generated images and compare the resulting judgments with those produced by three VLM evaluators. In total, we collect 2,304 binary human judgments over 192 generated frames from eight stories and six representative generation methods. Each frame is assessed using three binary questions corresponding to the character, object, and background state dimensions.

Human annotators and VLM evaluators follow the same predefined rubric. A "Yes" indicates that the depicted state agrees with the corresponding ground-truth state, whereas a "No" indicates a mismatch. Table 8 summarizes the shared evaluation rubric.

We use mean binary scores to aggregate human and VLM judgments. For each model and entity type, human judgments are averaged across annotators as the proportion of "Yes" responses, and VLM judgments are aggregated analogously. We then compute Spearman correlation between the human- and VLM-aggregated scores for each entity type.

Table 8: **Shared rubric for human and VLM evaluator validation.** Each visual-state dimension is assessed using one binary Yes/No question.

Entity Type	#	Yes/No Question
Character	1	*Does the character match with the state described in the provided ground truth?*
Object	1	*Does the object match with the state described in the provided ground truth?*
Background	1	*Does the background match with the state described in the provided ground truth?*
Total	**3**	

Table 9: **Spearman correlation between automated VLM judges and human evaluations across six representative methods, reported by visual property.** Higher values indicate stronger agreement with human judgments. **Bold** represents the highest score.

Visual Property	Gemini	Gemma	Qwen
Character	0.812	**0.829**	0.324
Object	**0.943**	0.899	0.886
Background	**0.986**	0.638	0.899
Average	**0.914**	0.789	0.703

Table 9 reports the Spearman correlation between each VLM evaluator and the corresponding human judgments across visual properties. While Gemini achieves the highest average correlation, we use `Gemma-4-31B-it` for the full benchmark. Gemma maintains strong agreement with human judgments while being sufficiently lightweight for self-hosted inference, enabling large-scale evaluation without recurring costs.

COI Evaluator Validation. We validate the VLM-based interaction evaluator against PViC (Zhang et al., 2023) on frames containing annotated character–object interactions, as shown in Table 10. We consider four vocabulary configurations. *Story-derived* constrains the VLM to labels extracted from the story texts, *PViC* uses the original PViC vocabulary, *Open-ended* allows the VLM to produce object labels without a predefined vocabulary, and *Excluded* omits the object vocabulary from the evaluator input. When both object and verb vocabularies are restricted to the PViC label space, the VLM achieves 98.0% agreement with PViC ($\kappa = 0.902$), indicating that the evaluator closely reproduces PViC judgments under the same vocabulary constraints. We therefore use the VLM evaluator with story-derived vocabularies for the primary interaction metrics, as the PViC vocabulary covers only 170 of 1,433 interaction-bearing ground-truth frames ($\sim$12%) and is too restrictive for the broader interactions present in our stories.

Table 10: **Agreement between the VLM interaction evaluator and PViC under different object and verb vocabulary constraints.**

Object Vocabulary	Verb Vocabulary	Agreement	κ
Story-derived	Story-derived	0.794	0.287
PViC	PViC	**0.980**	**0.902**
Open-ended	PViC	0.770	0.346
Excluded	PViC	0.715	0.411

C.2 VLM-Based Evaluation Pipeline

Evaluator Configuration. We use `Gemma-4-31B-it` as the common VLM evaluator for visual-state extraction, interaction recognition, and semantic judgments across all evaluated models. Structured outputs are decoded into predefined schemas and parsed programmatically. Binary evaluator judgments are obtained from three independent evaluations at a sampling temperature of 0.7 and aggregated using majority voting. The same evaluator configuration, prompts, and matching procedures are used for every evaluated model.

Visual State Graph Extraction. For each generated frame F_t, the evaluator extracts the visually observable characters, objects, and background states using the same representation as the corresponding ground-truth state graph. The evaluator receives the target frame together with benchmark-side identity references and canonical entity labels for identification, while all state values are inferred solely from the target frame, without access to the story text or temporal history.

The extracted state is represented as $\hat{G}_t = (\hat{C}_t, \hat{O}_t, \hat{B}_t)$, where $\hat{C}_t$, $\hat{O}_t$, and $\hat{B}_t$ denote the predicted character, object, and background states, respectively. The same extracted representation is used for entity-count, visual-state, and state-transition evaluation. For character counting, only the canonical benchmark characters associated with reference images are retained.

The following system prompt provides a definition of the procedure for extracting the visual state graph and restricts the model to reporting only information that can be observed visually from the target frame. The reference portraits used in this extraction step serve only to identify listed characters in the target frame and are independent of the character identity metric defined in Section 3.1.

System Prompt for Visual State Graph Extraction

You are extracting a visual state graph. You receive identity-reference portraits followed by one target story frame.

{`reference_lines`}
– Image {`target`} is the **ONLY** target frame to annotate.
Images 1 through {`n_refs`} are identity references. Image {`target`} is the **ONLY** target frame to annotate.

Use the reference portraits as identity evidence: every allowed character has one. Anyone else in the frame – a crowd, guards, a bystander, an unnamed extra – is **NOT** a character here. Do not report them and do not count them, however many the frame shows. Derive every state from Image {`target`} only. Never copy injury, clothing, held items, location, or condition from a reference portrait. Do not use or infer story text, prompts, history, thoughts, intentions, hidden conditions, or offscreen facts.

Rules:

- Allowed character names: {`character_names`}.
- Include a character only when that reference character is visibly present in the target. Omit absent characters completely; never write absent/offscreen.
- Allowed object names: {`object_names`}. Include an object **ONLY** when it is visibly present in the target frame, and write its name **EXACTLY** as listed. Omit the rest. Never report an object outside this list, and never rename one.
- Describe each entity with **SLOTS**, not one blended phrase. For every entity you report, fill **EVERY** slot the frame lets you judge – a slot left out is read as "this image does not show it", so leave one

out only when the frame genuinely cannot show it. Use "none" only for held, when the hands are visibly empty; never write "unknown", "not visible" or any other non-answer as a value:

- **characters:** posture, outfit, injury, held
  posture = body pose or action;
  outfit = clothing and worn gear: when the character wears one of their outfits listed in {outfit_names}, write it **EXACTLY**; otherwise describe what they wear in one to three words;
  injury = "yes" if a wound, blood, bruise, burn or frostbite is visible **ON THE BODY**, otherwise "no" (dirt, sweat, flushed or pale skin, torn clothes and blood only on clothing are "no");
  held = what the hands hold: when it is one of {object_names}, write that name **EXACTLY**; "none" for empty hands; otherwise name what the hands hold in one to three words.
- **objects:** condition, location
  condition = physical state; location = where it sits or what it attaches to.
- **background:** place, weather, time
  place = setting: when the frame shows one of this story's settings, {place_names}, write it **EXACTLY**; otherwise name the setting in one to three words;
  weather = conditions; time = time of day. Omit weather or time when the picture cannot settle it – never write "none" for them.

- Each slot value is one to three words. Never put two slots in one value.
- "none" is an observation about the hands (they hold nothing), not a way to skip a slot. Omit a slot only when the frame cannot show it at all.
- A body part is never a character and never an object. Report it as a slot on the person it belongs to (a frostbitten hand makes the character's injury "yes").
- Write slot values in {language}. Keep canonical character and object names exactly as supplied. A value chosen from a list above, and the words "yes", "no" and "none", are copied exactly and never shortened or translated.
- Return strict JSON only, with no markdown or explanation.

Required JSON shape:

```
{"characters": [{"name": "canonical name",
   "slots": {"posture": "1-3 words", "outfit":
"listed outfit",
       "injury": "yes|no", "held": "listed object"}}],
  "objects": [{"name": "visible object",
    "slots": {"condition": "1-3 words", "location":
"1-3 words"}}],
  "background": {"slots": {"place": "listed setting",
       "weather": "1-3 words", "time": "1-3 words"}}}
```

Entity Alignment. Before comparing predicted and ground-truth states, we establish a one-to-one correspondence between entities in $\hat{G}_t$ and G_t. This accounts for generated frames in which entities may be missing, additional, or returned in a different order. Character entities are aligned according to their benchmark identities, while object entities may be matched semantically when different surface forms refer to the same object. Among compatible entities, we select a one-to-one assignment based on their observable state agreement. Unmatched reference entities are treated as missing predictions, while unmatched predicted entities are treated as additional predictions.

Semantic Matching. Open-form descriptions may express the same observable property using different lexical forms. We therefore use the VLM evaluator to determine whether two open-form state values or interaction descriptions are semantically equivalent. The decision is binary, with equivalent descriptions counted as matches and descriptions that alter, omit, or introduce an observable property counted as mismatches. Closed-valued properties are normalized and compared directly.

Character–Object Interaction Extraction. We extract visible character–object interactions from each target frame to evaluate whether the generated frames correctly capture the interactions specified by the story. During COI extraction, the evaluator is provided with the canonical character names and their corresponding reference images. Only interactions involving these listed characters are retained; any other visible person is ignored. Interactions are determined solely from visible physical contact or manipulation in the target frame.

System Prompt for Character–Object Interaction Extraction

Images 1 through {n_refs} are identity reference portraits.
{reference_lines}
Image {target} is the story frame to inspect. Use the portraits **ONLY** to decide which listed character a person in the frame is; read every interaction off Image {target}. A person in the frame who matches no portrait is not a listed character: ignore them.

List every visible human-object interaction. A subject must be a character entity and an object must be an object entity from the roster. Name each interaction with a concise {verb_language} base-form verb that describes the specific action actually shown in the image.

A valid interaction must be **PHYSICAL**: the character visibly contacts, holds, carries, wears, wields, pushes, pulls, moves, or otherwise physically acts on the object. Physical contact or physical manipulation must be visibly supported by the target image.

Do not infer interactions that are not visibly depicted. Do not return character-character, object-object, pose-only, or proximity-only relations – sharing a frame with an object is not attention, and attention needs the person's gaze to be legibly directed at that object.

If no listed roster entities form a visible valid interaction, return exactly `[]` without an explanation, apology, or refusal.

Entity roster:

{roster}

Return ONLY a JSON list:

```
[{"subject_id": "entity id", "verb": "base verb",
"object_id": "entity id"}]
```

C.3 Visual Quality Metrics

Text–Image Alignment. For each frame F_t, UnifiedReward (Wang et al., 2025c) is given the image, the fixed question "Describe this image accurately and comprehensively.", and the corresponding narrative sentence x_t as the response to be evaluated. Let $U(x_t, F_t) \in [0, 100]$ denote the resulting score. We normalize each frame-level score by 100 before averaging across the story:

$$\mathrm{UR} = \frac{1}{T} \sum_{t=1}^{T} \frac{U(x_t, F_t)}{100}.$$

Aesthetic Quality. We use a LAION aesthetic predictor (Schuhmann et al., 2022), implemented as an MLP head over L2-normalized CLIP ViT-L/14 image embeddings (Radford et al., 2021). Let $a(F_t)$ denote its raw aesthetic score for frame F_t. We first average the raw scores across the story and then normalize the resulting story-level score by 10:

$$\mathrm{AES} = \frac{1}{10T} \sum_{t=1}^{T} a(F_t).$$

D FULL RESULTS

We performed a benchmarking analysis of 21 text-to-image and story visualization models on STORYDYNAMICS, considering metrics related to characters, objects, interactions, state transitions, and visual quality. Table 11 presents a full comparison of the models and their generation paradigms.

We further analyze performance at the visual-property level to identify which character, object, and background properties are more reliably depicted and updated over time. Table 12 reports the mean performance across all 21 evaluated models.

We also analyze how the story settings affect model performance. Table 13 reports results across story length, language, sentence complexity, and visual style, averaged over the 21 evaluated models.

Table 11: **A full comparison of 21 text-to-image and story visualization models.** We report character and object count accuracy, character identity preservation (ID), character–object interaction F1 and exact match (EM), state and transition F1 for characters, objects, and backgrounds, UnifiedReward (UR), aesthetic quality (AES), and the **Overall harmonic mean**, as defined in Section 3.1. Higher is better. **Bold** and underlined values indicate the best and second-best results, respectively. Models are grouped by model family and generation paradigm.

Model	Count		ID	COI		State			Trans.			Quality		Overall ↑
						StateF1 ↑			TransF1 ↑					
	Cnt_{char} ↑	Cnt_{obj} ↑	ID ↑	$F1_{COI}$ ↑	EM_{COI} ↑	Char.	Obj.	Bg.	Char.	Obj.	Bg.	UR ↑	AES ↑	
Text-to-Image Models														
Generative Models														
GLM-Image	0.439	0.336	**0.864**	0.089	0.038	0.236	0.162	0.382	0.359	0.160	0.290	0.064	0.639	0.131
GPT-Image-1-mini	0.419	0.410	0.713	0.103	0.035	0.279	0.311	**0.578**	**0.497**	0.336	**0.462**	0.101	0.618	0.148
OmniGen2	0.625	0.445	0.511	0.116	**0.057**	0.266	0.220	0.359	0.371	0.252	0.310	0.087	0.625	0.173
Qwen-Image-Edit	0.573	0.459	0.392	0.094	0.042	0.259	0.294	0.558	0.403	**0.354**	0.414	0.136	0.620	0.162
Flux.1	0.638	0.266	0.040	0.033	0.011	0.256	0.107	0.319	0.289	0.108	0.240	0.053	0.671	0.050
Flux.2	**0.695**	**0.486**	0.767	**0.134**	0.048	**0.348**	**0.324**	0.468	0.443	0.350	0.343	**0.151**	0.635	**0.193**
Stable Diffusion XL (SDXL)	0.453	0.218	0.017	0.020	0.010	0.140	0.059	0.317	0.202	0.094	0.255	0.046	0.637	0.035
Stable Diffusion 3.5	0.607	0.277	0.039	0.035	0.015	0.233	0.128	0.343	0.260	0.139	0.266	0.065	0.651	0.059
Training-Free Models														
CharaConsist	0.447	0.223	0.026	0.037	0.018	0.198	0.073	0.399	0.262	0.074	0.257	0.045	**0.688**	0.055
Trained Models														
UNO	0.557	0.264	0.355	0.047	0.021	0.206	0.090	0.287	0.238	0.089	0.227	0.044	0.641	0.080
MS-Diffusion	0.380	0.254	0.001	0.003	0.000	0.074	0.087	0.242	0.079	0.079	0.223	0.051	0.619	0.000
Story Visualization Models														
Training-Free Models														
StoryDiffusion	0.468	0.180	0.010	0.017	0.014	0.132	0.046	0.351	0.174	0.042	0.230	0.033	0.652	0.032
Story Iter	0.392	0.175	0.002	0.004	0.002	0.084	0.040	0.289	0.121	0.028	0.218	0.032	0.601	0.007
ConsiStory	0.450	0.186	0.001	0.001	0.000	0.132	0.008	0.252	0.114	0.005	0.177	0.015	0.575	0.001
One-prompt-one-story	0.366	0.248	0.002	0.003	0.000	0.074	0.062	0.312	0.089	0.069	0.252	0.042	0.588	0.002
Trained Models														
StoryGen	0.380	0.199	0.023	0.001	0.001	0.084	0.033	0.263	0.086	0.016	0.239	0.012	0.503	0.003
SEED-Story	0.286	0.163	0.000	0.000	0.000	0.016	0.011	0.299	0.017	0.001	0.295	0.006	0.504	0.000
Agentic Models														
Anim-Director	0.642	0.253	0.002	0.011	0.007	0.247	0.106	0.337	0.178	0.090	0.274	0.062	0.636	0.012
MMStoryAgent	0.481	0.229	0.000	0.010	0.005	0.103	0.052	0.253	0.096	0.057	0.201	0.045	0.634	0.000
MovieAgent	0.581	0.245	0.005	0.018	0.012	0.201	0.082	0.304	0.177	0.089	0.245	0.059	0.635	0.024
Video Models														
Vlogger	0.393	0.250	0.005	0.003	0.001	0.093	0.066	0.226	0.075	0.077	0.206	0.040	0.478	0.006

Table 12: **Visual-property-level performance averaged across all 21 evaluated models.** $\mathrm{StateF1}$ measures whether the expected visual property is depicted correctly, while $\mathrm{TransF1}$ measures whether changes in that property occur correctly over time. Higher is better.

Entity	**Visual Property**	$\mathrm{StateF1}\uparrow$	$\mathrm{TransF1}\uparrow$
Character	Posture	0.123	0.205
	Outfit	0.047	0.353
	Injury	0.147	0.358
	Held	0.194	0.237
Object	Condition	0.125	0.139
	Location	0.098	0.094
Background	Place	0.426	0.292
	Weather	0.194	0.217
	Time	0.162	0.195

Table 13: **Performance across different story settings, averaged over the 21 evaluated models.** n denotes the number of stories.

Setting	n	**Count**		**ID**	**COI**	**State**	**Trans.**
		$\mathrm{Cnt}_{\mathrm{char}}\uparrow$	$\mathrm{Cnt}_{\mathrm{obj}}\uparrow$	$\mathrm{ID}\uparrow$	$\mathrm{F1}_{\mathrm{COI}}\uparrow$	$\mathrm{StateF1}\uparrow$	$\mathrm{TransF1}\uparrow$
Short	48	0.533	0.248	0.228	0.050	0.242	0.218
Long	45	0.442	0.303	0.177	0.023	0.174	0.188
English	45	0.532	0.325	0.170	0.047	0.233	0.234
Chinese	48	0.449	0.227	0.189	0.027	0.186	0.168
Simple	37	0.495	0.356	0.185	0.034	0.206	0.199
Complex	56	0.485	0.221	0.177	0.039	0.211	0.203
Anime	20	0.397	0.266	0.167	0.026	0.236	0.205
CGI	29	0.537	0.232	0.210	0.033	0.195	0.184
Claymation	18	0.505	0.269	0.200	0.038	0.216	0.189
Photorealism	26	0.495	0.333	0.150	0.048	0.198	0.221

E VALIDATION AGAINST A PROFESSIONAL CINEMATIC NARRATIVE

To provide further detail on the cinematic validation described in the main text, we apply the full STORYDYNAMICS evaluation pipeline to *One Battle After Another* (Figure 3). A VLM (`gpt-5.5-2026-04-23`) first derives a textual storyline from the film sequence, which is subsequently converted into the structured reference annotations used by STORYDYNAMICS. The original film frames are then treated as predictions and evaluated against these references using the same metrics and evaluation procedure as in the benchmark experiments. Table 1 reports the resulting dimension-wise and overall narrative scores.

We further examine the visual-state scores at the property level for a representative segment of the film. Table 14 reports the corresponding character, object, and background properties.

To further examine the metric behavior across entity types, Table 15 reports the corresponding StateF1 and TransF1 scores for characters, objects, and backgrounds, together with the number of applicable transition edges.

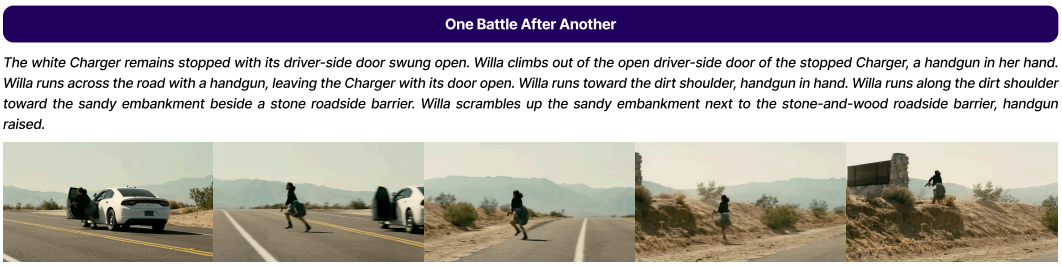


Figure 3: Application of STORYDYNAMICS to *One Battle After Another*. A textual storyline and structured reference annotations are derived from the film, while the original frames are evaluated using the same pipeline as the benchmark experiments.

Table 14: **Visual-property-level state performance on *One Battle After Another*.** StateF1 measures agreement between the predicted and reference visual states for each property. GT States denotes the number of applicable ground-truth states. A dash indicates that no applicable ground-truth state exists.

Entity	Visual Property	StateF1 ↑	GT States
Character	Posture	0.800	5
	Outfit	1.000	5
	Injury	–	0
	Held	0.800	5
Object	Condition	1.000	2
	Location	0.857	8
Background	Place	1.000	6
	Weather	–	0
	Time	–	0

Table 15: **Entity-level visual-state and state-transition performance on *One Battle After Another*.** $\mathrm{StateF1}$ measures agreement with the reference visual states, while $\mathrm{TransF1}$ measures whether the annotated state transitions are correctly realized. A dash indicates that no applicable ground-truth transition exists.

Entity	$\mathrm{StateF1}\uparrow$	$\mathrm{TransF1}\uparrow$	**Transition Edges**	
			Reached	**Total**
Character	0.867	1.000	2	2
Object	0.889	–	0	0
Background	1.000	–	0	0
Overall	**0.919**	**1.000**	**2**	**2**

F STORY VISUALIZATION

We provide visualization results that represent different story settings, including length, complexity and language. For each combination, we select a single story visualization model.

F.1 SHORT STORY

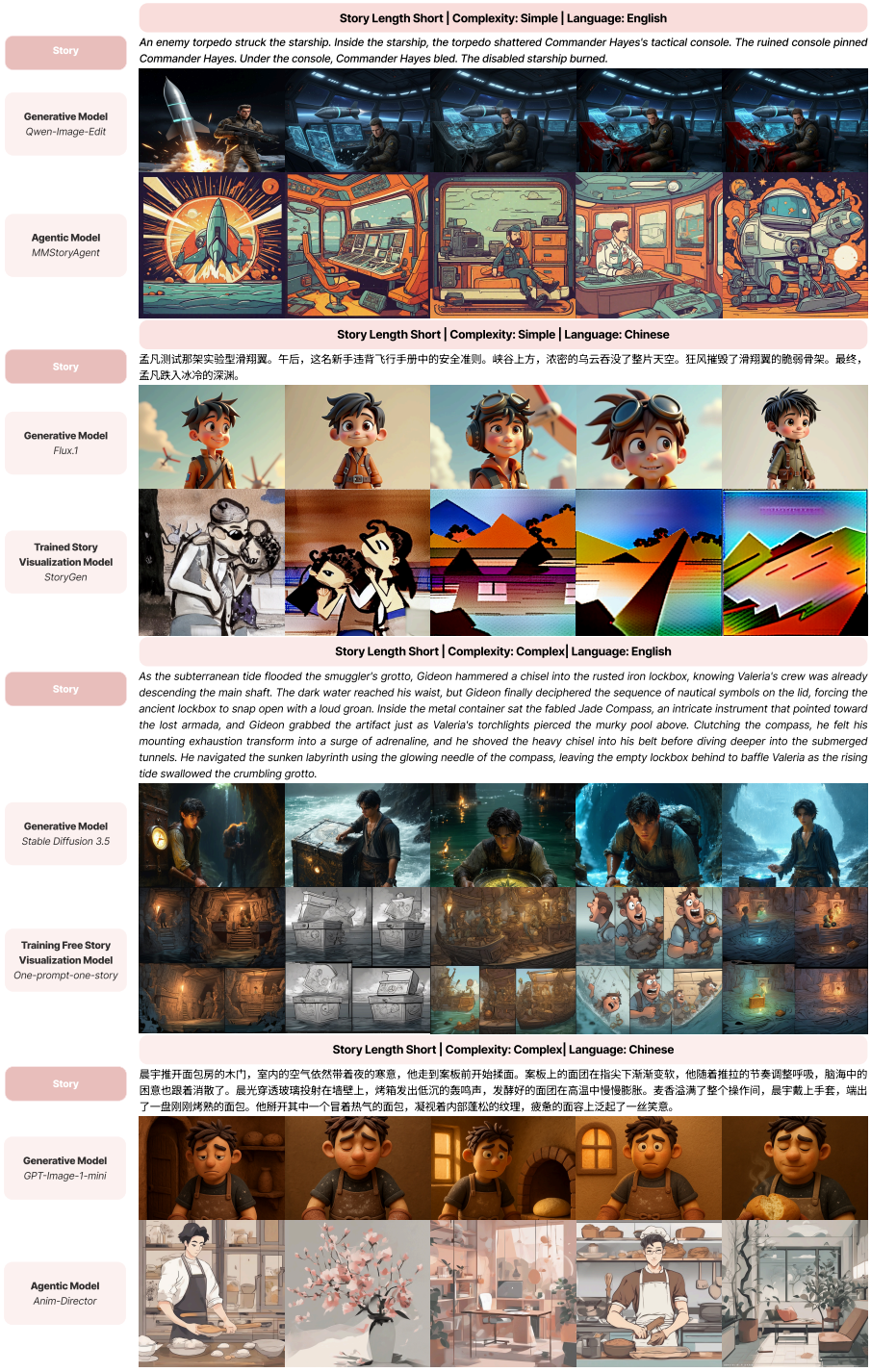


Figure 4: Short-story visualizations across representative model families.

F.2 LONG STORY

Story Length Long| Complexity: Simple | Language: English

Story

Niles stood inside the stone fortress. The colonial harbor stretched below him. Mercer loaded the bronze cannon. An enemy fleet blockaded the port. The wooden ships flew imperial colors. Niles gripped the wooden ramrod. The wooden ramrod pushed the gunpowder. The bronze cannon felt cold. Mercer primed the touchhole. The enemy frigates sailed closer. Niles withdrew the wooden ramrod. He stepped away from the bronze cannon. Mercer lit the fuse. The bronze cannon fired a heavy iron ball. The recoil shook the stone floor. Smoke filled the enclosed battery. Niles coughed. The iron ball struck a distant frigate. Wood splintered on the enemy deck. Mercer cheered at the direct hit. Niles grabbed the wooden ramrod again. The enemy ships returned fire. Cannonballs smashed into the fortress wall. Dust rained onto the bronze cannon. Niles attached a wet sponge to the wooden ramrod. He extinguished the lingering embers. Mercer brought another powder charge. Niles guided the wooden ramrod. He pushed the new iron projectile inside. The bronze cannon grew warm. The afternoon sun began its descent. Shadows lengthened across the colonial harbor. The enemy fleet maintained the blockade. Mercer ignited the second fuse. The bronze cannon roared. The wooden ramrod vibrated in Niles' hands. Another enemy ship sustained damage. The naval warfare escalated. The frigates unleashed a coordinated broadside. Stone debris showered the defenders. Niles ducked behind the bronze cannon. Mercer prepared the next shot. The sun dipped below the horizon. Twilight enveloped the colonial port. The fading light obscured the enemy vessels. Niles squinted into the gathering darkness. He relied on the muzzle flashes. Mercer yelled a new firing coordinate. Niles plunged the wooden ramrod into the barrel. The bronze cannon radiated heat. Niles rushed the loading process. The darkness demanded faster volleys. Mercer touched the match to the powder. The bronze cannon blasted a fiery plume. Night conquered the sky. The dark harbor swallowed the enemy fleet. Niles felt a strange vibration. He inspected the bronze cannon. A hairline fracture appeared near the breech. The metal suffered from the constant firing. Mercer ignored the structural warning. He poured more powder into the touchhole. Niles hesitated with the wooden ramrod. The enemy cannons flashed in the blackness. The fortress took another direct hit. Niles shoved the roundshot into the barrel. The bronze cannon hissed. The metal crack widened by an inch. Mercer applied the glowing match. The bronze cannon detonated with lethal force. The breech ruptured in half. Hot metal fragments flew through the air. A shockwave knocked Niles backward. He dropped the wooden ramrod. The shattered bronze cannon smoked. Niles hit the stone wall. A ringing noise filled his ears. He lost all sense of hearing. Mercer mouthed words in the silence. Niles stared at his blistered palms. The hot metal had seared his skin. He touched his bleeding ears. Total deafness isolated him. The naval battle raged outside in silence. Niles crawled toward the ruined bronze cannon. He retrieved the discarded wooden ramrod. The useless stick offered no protection. Mercer dragged Niles away from the embrasure. The deafened gunner trembled. The enemy fleet bombarded the dark fortress. Niles felt the vibrations through the floor. He could not hear the explosions. His burned hands throbbed. The destroyed bronze cannon blocked the doorway. Mercer pointed toward the inner courtyard. Niles leaned on the wooden ramrod. He stood up on unsteady legs. The colonial defense line had fallen. The two men retreated into the shadows. They abandoned the smoking fortress.

Generative Model
GLM-Image

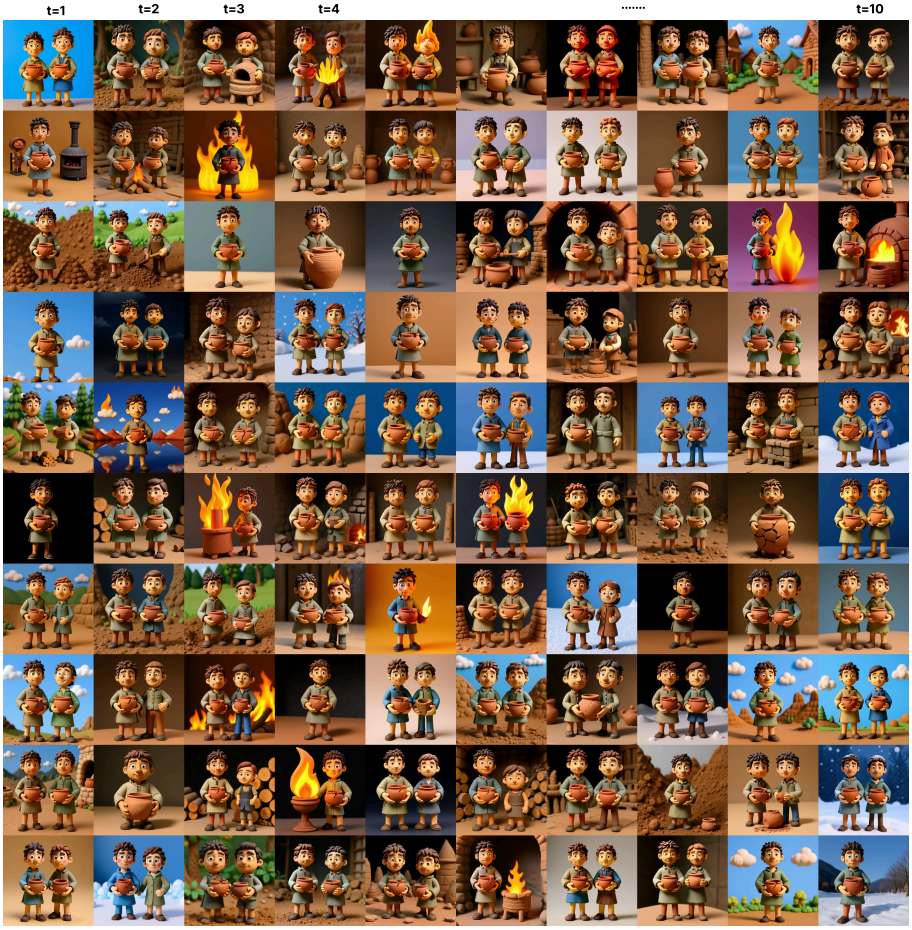


Story Length Long| Complexity: Simple | Language: Chinese

Story

天空呈现出蓝色。景明揉捏着泥土。窑炉立在旁边。里面燃烧着木柴。火焰温暖着景明。他烧制了一个陶罐。陶罐散发着红光。景明喜欢这个陶罐。弘文来到了这里。沉重的泥土伴随着弘文。远处立着一个大窑炉。弘文的窑炉吞噬着木柴。刺眼的火焰照亮了四周。弘文也烧制了一个陶罐。那个新陶罐非常耀眼。景明盯着弘文的陶罐。两个陶罐截然不同。景明讨厌自己的陶罐。弘文的陶罐刺痛了景明。景明摔碎了自己的陶罐。泥土堆满了四周。景明挖出了大量泥土。他模仿着弘文。巨大的新陶罐诞生了。天空变得灰暗。景明打开窑炉。窑炉内部十分空旷。景明添加了许多木柴。火焰逐渐升高。窑炉承受着火焰。景明观察着天空。天空变得阴暗。寒风吹过窑炉。景明躲避着寒风。完美的陶罐诱惑着景明。景明再次看向弘文。弘文正在雕刻泥土。那个陶罐十分精美。景明更加嫉妒弘文。窑炉吞噬了所有木柴。景明继续寻找木柴。天空倒映着火焰。窑炉裂开了。景明无视了窑炉。景明只想要那个陶罐。寒风吹打着弘文。弘文收起了他的陶罐。寒风迎接着弘文。景明独自守着窑炉。寒风加剧了。天空完全变黑了。景明投入最后的木柴。火焰冲出窑炉。窑炉崩塌了。景明失去了窑炉。火焰慢慢熄灭。景明寻找他的陶罐。泥土散落一地。那个巨大的陶罐碎裂了。景明没有得到完美的陶罐。原来的陶罐也消失了。寒风吹拂着泥土。景明跪在泥土上。木柴烧光了。景明无法再次点燃火焰。窑炉变成了泥土。寒风冰冻了泥土。天空变得漆黑。景明回忆着弘文的陶罐。弘文拥有美丽的陶罐。景明仰望着天空。结实的陶罐曾经属于他。温暖的火焰也曾属于他。那个完好的窑炉不复存在。景明浪费了许多木柴。地上的泥土已经变得坚硬。景明拼凑着那个陶罐。寒风带走了火焰。景明凝视着天空。天空毫无生气。弘文的窑炉立在远处。那个美丽的陶罐完好无损。充足的木柴堆在远处。弘文控制着火焰。景明盲目地追逐弘文。过多的木柴毁掉了窑炉。景明逼迫着窑炉。泥土散落在四周。景明砸碎了最初的陶罐。寒风包围了景明。景明不停地颤抖着。冰凉的寒风刺痛了景明。景明抓起一把泥土。泥土掉落在地上。景明看着破碎的窑炉。窑炉无法再燃烧木柴。曾经的火焰彻底消失了。任何陶罐都无法诞生。景明独自面对着天空。寒风吞噬了景明。

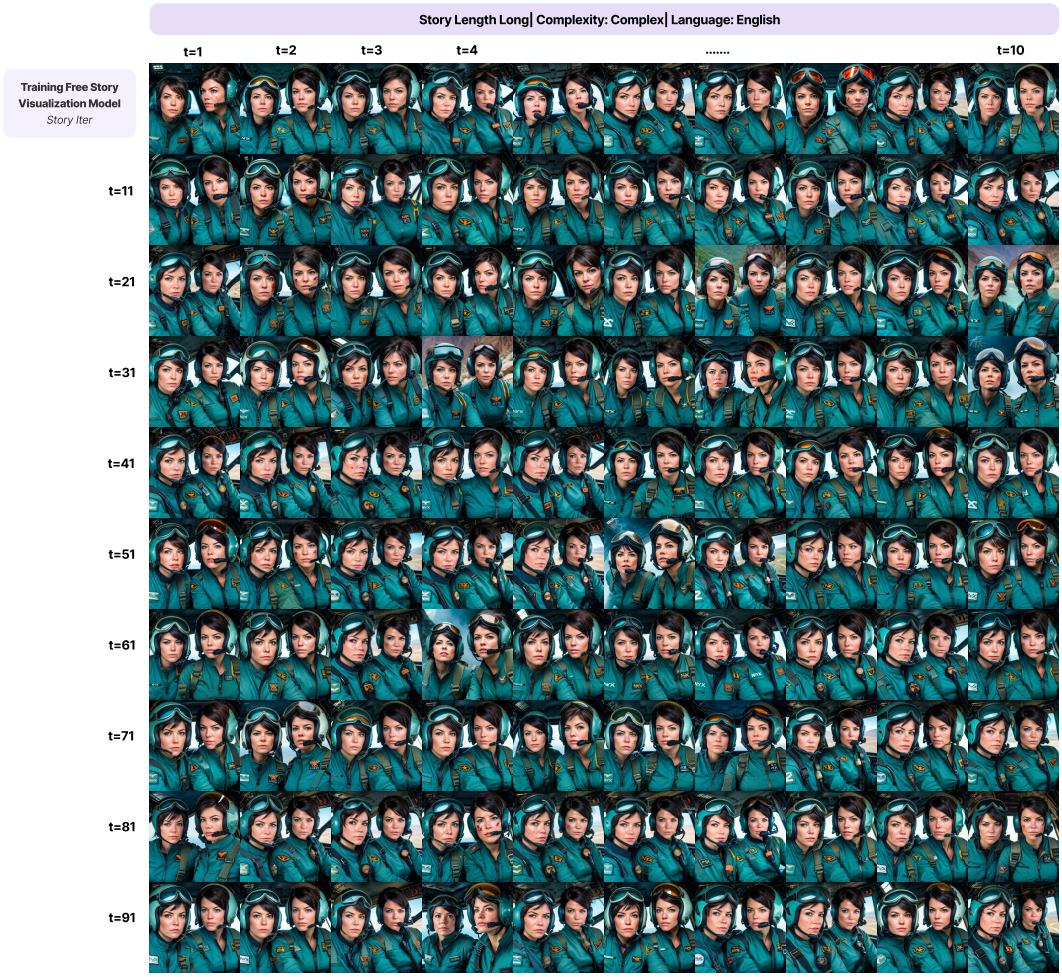
Generative Model
OmniGen2

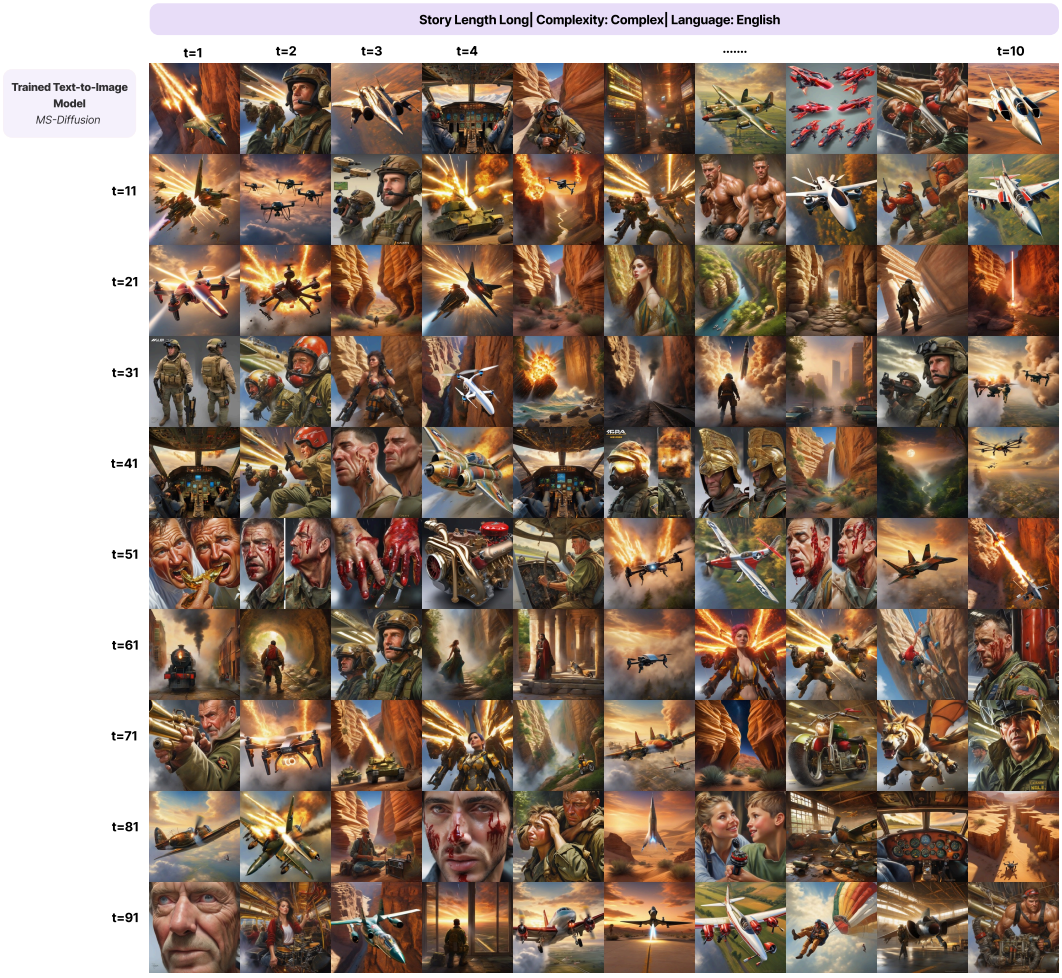


Story Length Long| Complexity: Complex| Language: English

Story

Galen ignited the afterburners, and his interceptor tore through the midnight stratosphere above the jagged canyon. The stars provided minimal illumination, so he relied on his tactical visor to scan the horizon. Nyx flew her own interceptor in a loose echelon formation, and she watched the airspace for any thermal anomalies. The high-altitude air remained freezing, but the cockpit heaters hummed steadily to keep the pilots warm. Galen tapped the side of his helmet, and he calibrated the tactical visor while the canyon yawned below them. A telemetry spike flashed across the interface, which prompted Nyx to break radio silence. She reported a cluster of unregistered signatures, and Galen adjusted his flight path to intercept the bogies. The tactical visor painted five hostile drones in glowing red outlines, which revealed their hypersonic approach vector. Galen pushed the throttle forward, and he felt the crushing weight of gravity press against his chest. The interceptor accelerated past Mach two, and the resulting sonic boom echoed across the empty desert. Nyx armed her primary cannons, and she dropped her interceptor into a steep dive to flank the incoming swarm. The drones broke their tight formation, and they scattered across the night sky as they initiated evasive maneuvers. Galen tracked the lead drone, and he let the tactical visor calculate the necessary deflection angle for a shot. He squeezed the trigger, and the cannons sent a stream of depleted uranium rounds tearing into the enemy machine. The drone exploded into a fireball, and burning shrapnel rained down toward the canyon floor. Nyx pursued two flanking targets, and she wove through the thin air while she locked her weapon systems. The G-forces mounted fast, which forced Galen to tense his abdominal muscles so he could maintain consciousness. A surviving drone snapped into a vertical climb, because it attempted to outrun the interceptors in the upper atmosphere. Galen hauled back on the control stick, and he matched the aggressive ascent while his tactical visor flickered. The altitude robbed the engines of oxygen, which caused the interceptor to shudder under the immense strain. He fired a short burst, and the rounds clipped the drone's propulsion unit before the machine stalled in midair. The crippled drone spun out of control, and it detonated harmlessly as Nyx called out the remaining targets. Three hostile signatures abruptly plunged toward the earth, because they sought refuge within the narrow desert canyon. Galen inverted his interceptor, and he dived straight down while he chased the fleeing enemies into the dark abyss. The canyon walls rushed past in a blur, and the tactical visor struggled to map the jagged terrain. Nyx followed closely behind, and she warned Galen that the narrow confines left no room for error. The drones darted through the winding ravines, and they used the natural rock formations to break the tracking locks. Galen gritted his teeth, and he fought the violent turbulence as wind currents whipped through the stone corridors. He angled the interceptor sideways, and he slipped between two massive pillars while the tactical visor projected the optimal path. One drone deployed a barrage of flares, which blinded the thermal sensors and flooded the canyon with harsh light. Galen switched his tactical visor to optical mode, and he relied on visual tracking to maintain his pursuit. The extreme maneuvers drained his stamina, and sweat began to pool beneath his heavy flight helmet. Nyx banked sharply to avoid a rock wall, and she fired her cannons at the trailing drone as she leveled out. Her rounds struck the canyon wall, and the debris showered the drone before the machine lost aerodynamic stability. The machine slammed into the cliffside, and it erupted into a massive explosion that shook the surrounding rock. Thick black smoke billowed from the wreckage, and the smog spread quickly to choke the narrow canyon passageways. Galen plunged his interceptor into the expanding smoke cloud, and he lost sight of the stars above. The background environment shifted from clear darkness to a suffocating smog, which reduced external visibility to absolute zero. He trusted the tactical visor to navigate the gloom, and he kept his eyes glued to the digital terrain overlay. The remaining two drones used the smoke as cover, and they turned sharply to launch a counterattack against the pilots. Proximity alarms blared inside the cockpit, which warned Galen that an enemy projectile was approaching fast. He yanked the control stick hard left, and he pulled nine Gs as the interceptor rolled away from the threat. The physical toll became agonizing, and Galen felt a sharp pain radiate through his neck muscles. A glancing hit struck the nose of his interceptor, and the impact sent a violent shockwave through the entire fuselage. Sparks showered the cockpit interior, and the tactical visor emitted a high-pitched whine before glitching in a cascade of errors. The digital display cracked down the center, which rendered the tactical visor useless in the dense smoke. Galen tore the broken visor from his helmet, and he tossed the dead hardware aside as he gasped for air. Without the augmented display, he had to fly by spatial memory alone, so he navigated the obscured canyon blind. The smoke thickened further, and the toxic haze filled the ravine until it blotted out the moonlight. Nyx transmitted a frantic warning over the radio, and she stated that the drones were circling back for another pass. Galen tasted copper in his mouth, because the sustained G-forces had ruptured a vessel in his nose. Blood trickled down his chin, which marked his transition from a calculated pilot to a battered survivor fighting physical collapse. He wiped the blood away with his glove, and he gripped the control stick tightly while the interceptor rattled. The smoke severely hampered the engine intakes, which caused the thrust output to fluctuate unpredictably. Galen listened to the pitch of the turbines, and he adjusted the throttle manually to prevent a catastrophic stall. A drone burst through the smoke screen, and it passed within feet of Galen's canopy as it unleashed a volley of fire. He kicked the rudder pedals, and he skidded the interceptor sideways to dodge the deadly stream of tracers. The sudden lateral G-force wrenched his spine, and the motion pulled another groan of pain from his bleeding lips. Nyx locked onto the aggressor, and she fired a heat-seeking missile that tracked the drone through the smog. The missile found its mark, and the warhead annihilated the drone before adding more fiery debris to the choked canyon. Only one hostile remained, but the dense smoke made it impossible to locate the machine visually. Galen breathed in ragged gasps, and he fought the encroaching tunnel vision that threatened to render him unconscious. The loss of his tactical visor left him vulnerable, which forced him to rely on Nyx's radio telemetry. Nyx called out heading adjustments, and she guided Galen through the winding ravine while the smoke obscured the walls. He followed her voice implicitly, and he banked the interceptor by raw instinct as the stone pillars rushed past unseen. The final drone throttled down, because it attempted to ambush Galen from a hovering position within the darkest smoke layer. Nyx detected the thermal bloom, and she screamed the coordinates over the comms before the drone could fire. Galen shoved the throttle to maximum, and he rocketed the interceptor upward to break the enemy's targeting lock. The vertical climb subjected his body to crushing gravity, and the pressure squeezed the breath from his aching lungs. Blood flowed freely from his nose now, and the fluid stained his flight suit as he fought to keep his eyes open. He blindly squeezed the cannon trigger, and he swept the nose of the interceptor across the estimated target zone. A series of metallic impacts echoed through the smoke, which indicated that his blind volley had struck the drone. The enemy machine plummeted past his canopy, and it trailed fire before it shattered against the canyon floor. Nyx confirmed the final kill, and she praised his accuracy while she steered her own interceptor toward higher altitude. Galen kept his interceptor in a steep climb, because he was desperate to escape the suffocating smoke that filled the ravine. The engines roared in protest, and they fought the lack of oxygen until the aircraft finally breached the smog layer. Clear midnight air washed over the canopy, which revealed the pristine stars that hung above the desert landscape. Galen leveled the interceptor, and he eased the throttle back to let the overworked engines cool down. The release of the G-forces hit him hard, and the sudden change left his limbs trembling as the adrenaline began to fade. He looked at the discarded tactical visor, and he noted the cracked glass that had nearly cost him his life. Nyx formed up on his right wing, and she inspected his battered fuselage while they cruised through the calm stratosphere. She noted a scorch mark near the cockpit, and she marveled that the enemy projectile had not breached the hull. Galen keyed his radio, and he thanked Nyx for the navigational assistance during his blind flight through the canyon. He wiped his face again, and he smeared the dried blood across his cheek as his breathing slowly stabilized. The physical exhaustion weighed on his shoulders, and the fatigue replaced the earlier rush of combat. Nyx transmitted the return coordinates, and she set a course for the hidden desert airbase located miles away. Galen adjusted his heading to match hers, and he kept a loose grip on the control stick as they flew. The dead tactical visor rattled on the floorboards, and the broken hardware served as a reminder of the chaotic dogfight below. He checked his fuel gauges, and he verified that they had enough propellant to make the journey home safely. The desert stretched endlessly beneath them, and the shadows hid the scarred canyon where the destroyed drones lay burning. Galen felt a dull ache behind his eyes, because the extreme gravitational pressure had left a lingering impact on his body. Nyx maintained absolute radio silence now, and she allowed the rhythmic hum of the engines to soothe their frayed nerves. The interceptor flew smoothly through the undisturbed air, and the steady flight left the violence of the canyon far behind. Galen stared at the horizon, and he watched the faint glow of the distant airbase lights pierce the darkness. He prepared the aircraft for descent, and he lowered the landing gear while the base control tower cleared their approach. The runway illuminated the desert floor, and the lights guided the two interceptors down toward the concrete strip. Galen pulled the throttle to idle, and he let the aircraft glide gracefully until the tires kissed the tarmac. He engaged the braking parachutes, and he felt the sudden deceleration press him gently against his harness. Nyx landed smoothly right behind him, and she rolled her interceptor toward the designated maintenance hangars. Galen powered down the engines, and he unbuckled his harness while he finally let his exhausted muscles rest.

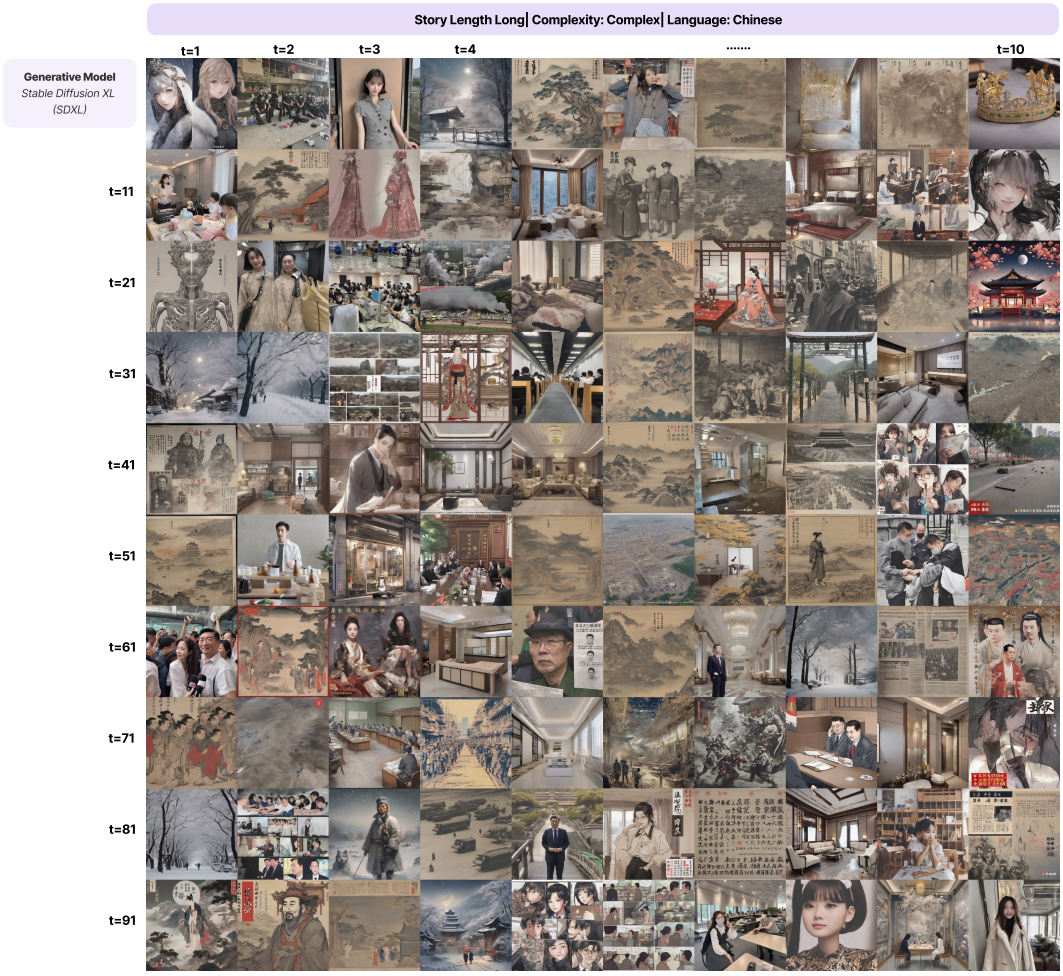

Story Length Long| Complexity: Complex| Language: English
t=1
t=2
t=3
t=4
.....
t=10
Training Free Story Visualization Model
Story Iter
t=11
t=21
t=31
t=41
t=51
t=61
t=71
t=81
t=91


Story Length Long| Complexity: Complex| Language: English
t=1
t=2
t=3
t=4
.....
t=10
Trained Text-to-Image Model
MS-Diffusion
t=11
t=21
t=31
t=41
t=51
t=61
t=71
t=81
t=91

Story Length Long| Complexity: Complex| Language: Chinese

Story

狂风裹挟着暴雪猛烈撞击律师事务所的窗户，谭睿盯着桌上的案卷，安雅则在一旁摆弄着那个氧化发黑的银质药盒。富商的猝死被警方初步判定为心脏病发作，但谭睿怀疑这是一场蓄意谋杀，安雅也觉得案卷里记录的物证存在疑点。长子赵阔声称父亲当晚只是按时服药，竞争对手却散布赵阔图谋遗产的谣言，谭睿必须在重重迷雾中梳理出真相。办公室里的温度因为暴雪引发的停电而不断下降，谭睿冻得双手发抖，他连翻阅案卷的动作都显得有些迟缓。安雅借着微弱的光线观察桌面，她试图徒手掰开那个生锈的银质药盒，但药盒的金属铰链卡得死死的。谭睿的心里布满了对即将到来的庭审的恐惧，他害怕自己无法击败赵阔那支昂贵的律师团队，这种焦虑让他不停地擦拭额头的冷汗。案卷里缺少直接指向赵阔的证据，谭睿只能将希望寄托在这个遗物上，他祈祷安雅能从银质药盒里找到突破口。窗外的积雪越来越厚，室内的气温急剧下降，银质药盒的金属外壳在低温下变得异常脆弱。安雅用双手死死扣住药盒的边缘，谭睿凑近观察她的动作，两人的呼吸在冰冷的空气中化作白雾。就在安雅用力下压边缘时，脆弱的铰链发出一声脆响，银质药盒的盖子断裂开来。断裂的盖子掉落在案卷上，药盒底部赫然显露出了一个隐蔽的夹层，谭睿的眼睛立刻瞪大了。夹层里藏着几粒红色的未知胶囊，安雅将它们小心地取出来，谭睿则立刻翻开案卷比对毒理学数据。死者生前服用的心脏病药物本应是白色的，而这些红色胶囊显然被掉包了，谭睿意识到这就是赵阔杀人的凶器。找到决定性证据后，谭睿内心的恐惧随之消散，他挺直了原本佝偻的脊背。暴雪依然在窗外肆虐，谭睿的眼神却变得无比坚毅，他指挥安雅立刻联系相熟的化验室进行成分分析。安雅冒着严寒跑出事务所去传递样本，谭睿将断裂的银质药盒收好，他已经构思好了在法庭上反击的策略。赵阔此前在媒体面前表现得极其悲痛，谭睿发誓要撕下这个伪君子的面具，安雅也对即将到来的战斗充满期待。化验室很快传回了加急报告，红色胶囊里装满了足以诱发心梗的烈性毒药，谭睿将报告单夹进了案卷里。案情的核心逻辑已经补齐，但谭睿还需要一名证人来证明赵阔接触过银质药盒，安雅立刻想到了赵家的保姆李玉。李玉在案发后便辞职回了老家，谭睿决定派安雅连夜去寻找她，他自己则留在办公室里继续完善案卷的细节。办公室的备用电源终于启动，谭睿在台灯下奋笔疾书，安雅则顶着风雪踏上了前往邻市的路途。那个断裂的银质药盒静静地躺在桌角，它虽然失去了完整的外观，却承载着揭开真相的全部重量。谭睿反复推敲着赵阔可能提出的辩护理由，他将每一条反驳意见都写在案卷的空白处，安雅在雪地里艰难前行。赵阔的律师团队以擅长伪造证据闻名，谭睿必须确保每一个环节都严丝合缝，他绝不允许这场官司出现任何纰漏。几个小时后，安雅在邻市的一处破旧出租屋里找到了李玉，李玉显得十分惊慌，她甚至不敢给安雅开门。安雅隔着门板向李玉描述了那个带有夹层的银质药盒，她解释了赵阔的罪行，李玉终于卸下防备打开了房门。李玉坦言自己曾在案发当晚看到赵阔在书房里摆弄银质药盒，但她因为害怕被报复而选择了沉默，安雅耐心地倾听着她的诉说。安雅向李玉保证警方会提供全天候的保护，李玉在犹豫了许久之后，终于答应出庭作证。谭睿接到安雅的确认消息后长舒了一口气，他合上面前的案卷，窗外的暴雪也开始逐渐减弱。漫长的一夜终于过去，谭睿在办公室的沙发上短暂地闭目养神，安雅则带着李玉踏上了返回市区的路程。清晨的第一缕阳光刺破了云层，持续了整晚的暴雪终于停歇，谭睿抱着厚重的案卷走出了律师事务所。积雪在阳光的照射下开始融化，街道上响起了扫雪车的轰鸣声，谭睿的步伐显得无比坚定。安雅直接将李玉送到了法院的地下车库，谭睿在那里与她们汇合，三人一起走楼梯前往等候室。那个断裂的银质药盒被谭睿装在口袋里，他能感受到金属传来的冰冷触感，这让他保持着绝对的清醒。法院大厅里挤满了前来报道的媒体记者，赵阔在保镖的簇拥下高调入场，谭睿则带着安雅从侧门进入了法庭。庭审正式开始，法官宣布全体肃静，谭睿从容不迫地走向了原告席。赵阔坐在被告席上露出不屑的冷笑，他的辩护律师率先站起身来，律师大声宣读着死者的自然死亡结论。辩方试图将案件定性为一场无理取闹的遗产纠纷，谭睿安静地听着对方的陈述，他没有表现出丝毫的慌乱。阳光透过法庭的高窗洒在木质地板上，室内的温度逐渐升高，谭睿翻开了面前的案卷。轮到原告方发言时，谭睿稳步走到法庭中央，他向陪审团展示了案卷里那份带有疑点的初步报告。谭睿指出死者体内的药物残留浓度极不合理，辩方律师立刻提出反对，法官则示意谭睿继续说明。谭睿传唤了第一位医学专家出庭，专家详细解释了毒药与心脏病药物的区别，赵阔的眉头微微皱了起来。辩方律师试图在交叉盘问中诱导专家推翻结论，谭睿及时提出抗议，法官当庭驳回了辩方的不当提问。安雅坐在旁听席的前排密切关注着局势，她将备用的案卷整理好，随时准备递交给谭睿。谭睿的每一次提问都直击要害，他不再是那个在办公室里发抖的年轻律师，法庭已经成了他掌控的领地。专家证人退场后，谭睿申请直接盘问被告，赵阔在律师的陪同下傲慢地走上了证人席。谭睿询问赵阔在案发当晚的行踪，赵阔坚称自己一直在卧室里睡觉，他试图用完美的不在场证明敷衍过去。阳光照在赵阔那张虚伪的脸上，谭睿冷冷地注视着他，安雅在台下握紧了拳头。谭睿抽出一份案卷里的监控记录，他指出监控录像在案发时段有两分钟的空白，赵阔的脸色微微变了。赵阔辩称那是监控系统的日常故障，谭睿却没有继续在这个问题上纠缠，他转身走向了原告席的物证台。谭睿拿起了那个银质药盒，断裂的盖子在空气中晃动，陪审团的目光全都被吸引了过去。他将银质药盒展示给赵阔看，他大声询问赵阔是否认得这件物品，赵阔强装镇定地点了点头。赵阔承认这是父亲随身携带的物品，但他强调这与案件无关，辩方律师也随声附和着提出抗议。谭睿没有理会辩方的干扰，他直接将断裂的银质药盒举高，安雅配合地大声宣读了化验室的毒理报告。谭睿当庭展示了银质药盒底部的隐蔽夹层，他向陪审团说明了夹层里藏有毒药的事实，法庭内响起一片惊呼。赵阔的额头上渗出了冷汗，他大声指责谭睿伪造证据，谭睿则将案卷里的化验单原件直接展示在众人面前。银质药盒的断口处闪烁着金属的光泽，谭睿逼问赵阔为何要定制这种带有夹层的容器，赵阔结结巴巴地无法回答。辩方律师急忙站起来解围，他声称银质药盒可能被任何人动过手脚，谭睿等的就是对方的这句话。谭睿顺势向法官提出申请，他要求传唤一位不在初始案卷名单上的关键人物，法官批准了他的请求。法庭的大门被推开，李玉在法警的护送下走入大厅，赵阔看到李玉的瞬间便瘫坐在了椅子上。阳光洒在李玉略显沧桑的脸庞上，她走到证人席前完成了宣誓，谭睿用温和的语气安抚着她的紧张情绪。李玉看了一眼旁听席上的安雅，安雅给了她一个鼓励的眼神，李玉终于鼓起勇气看向了陪审团。谭睿询问李玉在案发当晚看到了什么，李玉颤抖着声音描述了赵阔潜入书房的经过，法庭内安静得落针可闻。李玉作证说她亲眼看到赵阔扭开了银质药盒的夹层，他将几粒红色胶囊塞了进去，赵阔在被告席上发出了绝望的咆哮。法官严厉地警告了赵阔的扰乱行为，谭睿则继续引导李玉完善细节，那个断裂的银质药盒成为了证词的最佳佐证。辩方律师试图质疑李玉的视力问题，李玉却准确地描述出了银质药盒表面的繁复花纹，辩护律师哑口无言。谭睿将李玉的证词与案卷里的监控空白时段完美契合，他向陪审团梳理了赵阔完整的作案时间线，案件的真相已经大白于天下。阳光照亮了法庭的每一个角落，暴雪带来的阴霾被随风驱散，谭睿的陈述掷地有声。李玉完成了她的作证任务，她在法警的护送下退出了法庭，安雅悬着的心终于放了下来。谭睿拿着那个断裂的银质药盒走到陪审团面前，他发表了极具感染力的最后陈词，赵阔则绝望地低下了头。辩方律师放弃了最后的挣扎，他们知道这场官司已经毫无胜算，谭睿将银质药盒交还给书记员。法官宣布休庭等待陪审团商议结果，谭睿回到座位上整理散乱的案卷，安雅走过来对他表示祝贺。谭睿长长地呼出了一口气，他感到一阵畅快，那个曾经发抖的年轻律师已经蜕变成了真正的法庭斗士。赵阔被法警暂时带离了法庭，他路过谭睿身边时恶狠狠地瞪了一眼，但谭睿只是平静地回敬着他的目光。走廊里的阳光十分刺眼，安雅向谭睿汇报了外界媒体的轰动反应，这桩豪门谋杀案已经登上了新闻头条。休息时间很快结束，所有人重新回到法庭内落座，陪审团团长拿着裁决书站了起来。陪审团一致裁定赵阔的一级谋杀罪名成立，法庭内爆发出热烈的掌声，法官大声呼喊才维持住秩序。赵阔瘫软在椅子上无法动弹，法官当庭宣判了终身监禁的刑罚，谭睿将案卷重重地合上。那个断裂的银质药盒被正式封存进档案库，它完成了揭露罪恶的使命，安雅在案卷的封面上画下了一个圆满的句号。谭睿抱着案卷走下原告席，他与迎面走来的辩方律师握了握手，对方的眼神里充满了敬畏。走出法院的大门，冰冷的空气中弥漫着融雪的清新气息，谭睿深吸了一口气。媒体记者蜂拥而上试图采访这位一战成名的律师，安雅熟练地挡住了人群，她护送着谭睿走向路边。谭睿坐进汽车副驾驶座里闭上眼睛，他回想起昨晚在暴雪中撬开银质药盒的瞬间，心中感慨万千。安雅启动了汽车引擎，车辆平稳地驶入主干道，阳光在挡风玻璃上折射出耀眼的光斑。赵阔的商业帝国因为这场判决而迅速土崩瓦解，谭睿在车载广播里听到了相关的新闻，他默默地看向窗外。李玉已经安全地返回了老家，她向安雅发送了一条表达感谢的信息，安雅将内容转述给了谭睿。谭睿微笑着点了点头，他知道正义虽然充满波折，但案卷里的真相永远不会被掩埋在积雪之下。回到律师事务所后，办公室里的温度已经恢复了正常，谭睿脱下厚重的大衣挂在衣帽架上。安雅将赵阔案的案卷整理成册，她把那份毒理学报告放在了最上面，谭睿则站在窗前俯瞰着这座放晴的城市。曾经让他感到恐惧的法庭已经成为了他的主场，谭睿不再怀疑自己的能力，他的目光变得无比深邃。银质药盒的残骸虽然不在身边，但它断裂时的那声脆响依然回荡在谭睿的耳畔，这提醒着他时刻保持敏锐。事务所的老板亲自打来电话祝贺，谭睿礼貌地回应了几句，他并没有因为暂时的胜利而沾沾自喜。安雅倒了两杯热水端过来，她将其中一杯递给谭睿，两人默契地碰了碰杯子。阳光融化了窗台上的残雪，街道恢复了往日的喧闹，谭睿坐回办公桌前准备迎接新的工作。赵阔的案子正式画上了句号，谭睿将所有纸质案卷锁进保险柜，安雅则开始清理桌面上残留的杂物。法律的威严需要有人去捍卫，谭睿深知自己的职业道路还很漫长，他必须不断迎接新的挑战。办公桌上的通讯器突然响起，安雅接听后记录下新的委托信息，谭睿立刻拿起了面前的案卷。这是一起更为复杂的商业欺诈案，谭睿仔细聆听着安雅的汇报，他的大脑迅速进入了工作状态。阳光将两人的影子投射在背后的墙壁上，谭睿翻开了一本全新的空白案卷，安雅将新客户的资料递到了他的面前。无论未来的风雪有多么猛烈，谭睿都不会再有丝毫退缩，他带着坚定的信念投入了下一场战斗。


Story Length Long| Complexity: Complex| Language: Chinese
t=1
t=2
t=3
t=4
.....
t=10
Generative Model
Stable Diffusion XL (SDXL)
t=11
t=21
t=31
t=41
t=51
t=61
t=71
t=81
t=91


Story Length Long| Complexity: Complex| Language: Chinese
t=1
t=2
t=3
t=4
.....
t=10
Training Free Story Visualization Model
ConsiStory
t=11
t=21
t=31
t=41
t=51
t=61
t=71
t=81
t=91