DOI: 10.1111/cgf.70319

EUROGRAPHICS 2026 / B. Masia and J. Thies
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 45 (2026), Number 2

Story2Board: A Training-Free Approach for Expressive Visual Storytelling

D. Dinkevich[1], M. Levy[1], O. Avrahami[1], D. Samuel[2,3], and D. Lischinski[1]

[1]Hebrew University of Jerusalem, Israel

Figure 1: ***Story2Board*** *generates coherent multi-panel storyboards from a natural language prompt, maintaining subject identity while allowing dynamic changes in character pose, size, and position. Unlike prior work, it introduces a lightweight consistency mechanism that preserves the model's generative prior, supporting rich, expressive storytelling without fine-tuning or architectural changes.*

Abstract

We present Story2Board, a training-free framework for expressive storyboard generation from natural language. Existing methods narrowly focus on subject identity, overlooking key aspects of visual storytelling such as spatial composition, background evolution, and narrative pacing. To address this, we introduce a lightweight consistency framework composed of two components: Latent Panel Anchoring, which preserves a shared character reference across panels, and Reciprocal Attention Value Mixing, which softly blends visual features between token pairs with strong reciprocal attention. Together, these mechanisms enhance coherence without architectural changes or fine-tuning, enabling state-of-the-art diffusion models to generate visually diverse yet consistent storyboards. To structure generation, we use an off-the-shelf language model to convert free-form stories into grounded panel-level prompts. To evaluate, we propose the Rich Storyboard Benchmark, *a suite of open-domain narratives designed to assess layout diversity and background-grounded storytelling, in addition to consistency. We also introduce a new Scene Diversity metric that quantifies spatial and pose variation across storyboards. Our qualitative and quantitative results, as well as a user study, show that Story2Board produces more dynamic, coherent, and narratively engaging storyboards than existing baselines.* Project page: https://daviddinkevich.github.io/Story2Board/

CCS Concepts

• ***Imaging/Video*** → *Neural Image/Video Synthesis;* • ***Methods/Applications*** → *Artificial Intelligence/Machine Learning;*

1. Introduction

Text-to-image (T2I) diffusion models [HJA20; RDN*22; RBL*21; SCS*22; PEL*23] have rapidly transformed visual content creation, producing photorealistic and coherent images from natural language prompts with increasing reliability. Thanks to advances in open-source architectures and accelerated inference [RBL*21; SCS*22], these models have moved beyond research labs into creative workflows, such as illustrating children's books, powering social media campaigns, and supporting early-stage animation pipelines [LZL*23; ZMC*23]. As these models become more accessible, they are increasingly adopted not just as tools for static image generation, but also as engines for visual storytelling [YGL*24; HTC*25].

Storyboards represent a natural next step in visual storytelling. More than just sequences of snapshots, they are structured visual narratives, compositions that evolve across time, depicting characters, environments, and emotional beats in a spatially and semantically coherent manner. Effective visual storytelling relies not only on visual fidelity, but also on principles of cinematic composition: scale, perspective, framing, and environmental grounding [Blo20], as exemplified in Figure 1. Scenes such as a nomad dwarfed by a mount of bones, an empty beach under a stormy sky, or a girl in the glow of a treasure chest, communicate narrative meaning through spatial arrangement and atmosphere, rather than just subject appearance. Capturing this expressive diversity requires T2I models to move beyond static character rendering and embrace dynamic scene construction. This includes varying viewpoint and depth, emphasizing background storytelling, and adapting character presentation to reflect the evolving arc of the narrative [Fil25; Ani14].

Despite growing interest in automatic storyboard generation, current methods remain limited in their ability to produce visually compelling and narratively coherent image sequences. Several approaches focus narrowly on preserving character identity across frames, whether via reference-guided generation [TLY*24; YZL*23; WZJ*23], diffusion-based consistency [TKG*24b; HYT*24; ZZC*24], or autoregressive modeling [LWZ*24], but often do so at the expense of compositional diversity. As illustrated in Figure 2, generated characters are typically centered, scenes may lack spatial depth, and prompts tend to follow rigid templates such as "a photo of [character] in [setting]." As a result, these storyboards tend to resemble slideshows rather than expressive visual narratives.

To address these limitations, we propose a novel, training-free consistency framework that combines **Latent Panel Anchoring (LPA)** and **Reciprocal Attention Value Mixing (RAVM)** to guide modern T2I models toward generating coherent and expressive storyboards. **LPA** maintains a shared reference by anchoring the reference half of each two-panel latent across the batch during denoising, while allowing each scene-specific half to evolve independently. **RAVM** then reinforces character identity by mixing only the attention *value* vectors between reciprocally attended token pairs across panels, leaving queries/keys and attention routing unchanged. This values-only design preserves the model's compositional freedom (pose, layout, framing) while enforcing appearance consistency. Crucially, our method does not constrain the model's inherent generative capacity. Instead, it amplifies the in-context strengths of diffusion transformer (DiT) architectures by preserving a shared reference during denoising and softly blending appearance features between semantically aligned token pairs. This reinforces character identity and inter-panel coherence, while preserving the full compositional flexibility and visual richness of the base model. Importantly, our approach introduces no architectural changes or fine-tuning, offering token-level guidance that unlocks consistency without sacrificing diversity.

To interface with user input, we include a lightweight prompt decomposition step that converts natural-language stories into scene-level prompts using an off-the-shelf language model. This helps bridge freeform storytelling and visual generation, without requiring prompt engineering. The resulting method is compatible with state-of-the-art DiT-based models such as Stable Diffusion 3 and Flux [Sta24; Bla24], and examples of our outputs are shown in Figure 1.

While prior work emphasizes character consistency and prompt alignment, it largely overlooks a model's ability to convey story through composition and scene dynamics. Existing benchmarks [ZZC*24; HYT*24] rely on short, templated prompts with sparse environmental detail and limited narrative variation, so they rarely test whether models vary scale, pose, or position. We address this with the **Rich Storyboard Benchmark**, which is explicitly designed to elicit dynamic stories by requiring characters to change pose, scale, and placement across panels and to devote substantial description to backgrounds. Alongside it, we introduce two complementary evaluations designed to *quantify storyboard dynamism*: i.e., how character presentation changes across panels and how environments are depicted. **Scene Diversity** is our novel metric that aggregates changes in position, scale, pose, and visibility across panels, revealing the trade-off between identity consistency and narrative dynamics and exposing "consistency-by-repetition" failure modes. In parallel, a *background richness* user study scores environmental detail and grounding. Together, these measurements turn compositional change and environmental depiction into measurable quantities.

In summary, our contributions are threefold:

1. We introduce a novel training-free consistency framework that combines **Latent Panel Anchoring** and **Reciprocal Attention Value Mixing** to enhance in-context coherence in diffusion transformer models. This enables expressive storyboards with consistent characters, dynamic layouts, and rich environmental composition, without compromising diversity or requiring model fine-tuning.
2. We present the **Rich Storyboard Benchmark**, a suite of open-ended, visually grounded stories designed to evaluate layout flexibility, background detail, and narrative expressivity: dimensions underexplored in existing datasets.
3. We propose a new metric, **Scene Diversity**, which quantifies variation in character pose, scale, and framing across panels, offering a more nuanced assessment of visual storytelling beyond identity preservation.

Figure 2: *Comparative storyboard outputs from our method and two leading baselines, using the same input narrative. While baseline methods tend to center the character in every frame with limited variation in framing or environment, our method leverages cinematic principles such as exaggerated scale, dynamic perspective, and environmental context, to convey narrative progression more expressively. Note, for instance, how the small scale of the character in the third panel of the top row enhances the sense of vastness of the tower of bones, reinforcing the emotional arc of the story.*

2. Related Work

Text-to-image (T2I) diffusion models [HJA20; RDN*22; RBL*21; SCS*22; PEL*23] have revolutionized visual content generation, enabling high-quality synthesis from natural language prompts. Prominent recent models such as Flux [Bla24] and Stable Diffusion 3 [EKB*24] exemplify the capabilities of large-scale transformer-based [VSP*17] architectures in generating expressive, semantically grounded imagery. These models serve as a foundation for numerous methods for both consistent character synthesis and storyboard generation, two related but fundamentally distinct problem spaces.

Storyboard generation aims to produce sequences of images that together convey a narrative arc. The focus here is not solely on maintaining character identity, but on supporting dynamic compositions, evolving background elements, and expressive visual storytelling. In this space, StoryDiffusion [ZZC*24] introduces a consistency-aware attention module and a semantic motion predictor to guide narrative flow across frames. StoryGen [LWZ*24] introduces a learning-based autoregressive image generation model equipped with a vision-language context module, enabling coherent storyboard synthesis from freeform narrative input. DreamStory [HYT*24] similarly leverages a language model for prompt decomposition and employs a multi-subject diffusion architecture to preserve inter-character relationships across scenes. Other related efforts, such as IC-LoRA [HWW*24], explore lightweight adaptation techniques to improve generation coherence across time steps, while OminiControl [TLY*24] introduces image-based conditioning to guide spatial layout and stylistic coherence throughout a narrative.

In contrast, consistent character generation [AHV*24; TKG*24a] focuses on preserving the visual identity of a specific subject across multiple images. In this task, the character is typically the visual and semantic anchor of the composition, with background or narrative context playing a secondary role. Recent methods such as The Chosen One [AHV*24], ConsiStory [TKG*24a], and IP-Adapter [YZL*23] manipulate internal representations–either through iterative prompt-based refinement, cross-image feature sharing, or external adapters–to maintain consistency across scenes. While some of these methods include the term "story" in their titles or describe sequential results, their primary concern remains identity fidelity. For example, in ConsiStory [TKG*24a], consistency is measured almost exclusively through identity features, with little emphasis on narrative variation, layout dynamics, or background richness. This distinction is critical: consistent character generation centers the image around the character, whereas storyboard generation requires a broader representational range, where characters may appear small, partially occluded, shared with side actors, or absent altogether.

Our work targets storyboard generation, but differs in three ways. First, our pipeline is entirely training-free and applies directly to pre-trained transformer-based diffusion models such as Flux and Stable Diffusion 3 [EKB*24]. Second, it is character-agnostic: we require no character masks (e.g., SAM [KMR*23]) or reference tokens. Third, we avoid cross-image routing changes and instead mix only attention *values* between reciprocally attended token pairs (RAVM), improving identity while preserving composition.

Attention routing vs. value mixing. Many long-range consistency methods explicitly modify attention routing across frames (e.g., shared/extended attention or score edits) [HTC*25]. Such changes may transfer poorly to modern DiT-style backbones and can collapse layout variety [FHL*25]. In contrast, our **RAVM** preserves native routing (no attention-score, query, or key edits) and blends only *value* vectors for token pairs with strong reciprocal support.

3. Method

Our goal is to generate coherent storyboard panels from free-form text while preserving character identity across diverse compositions. Following [HYT*24; ZZC*24], we use an LLM (GPT-4o) to decompose the narrative into a shared reference panel prompt and n scene-specific prompts, which are then jointly rendered by a pre-trained diffusion model (see Fig. 3). To ensure consistency without retraining, we introduce two complementary mechanisms, *Latent Panel Anchoring (LPA)* and *Reciprocal Attention Value Mixing (RAVM)*. LPA pairs each panel with a shared reference, thereby leveraging the model's self-attention mechanism to promote visual consistency between panels (Section 3.1). While prompt-guided anchoring provides a useful bias, it alone is insufficient in scenes with complex layouts or ambiguous references. RAVM further enhances consistency by softly blending visual features between corresponding tokens across panels based on bidirectional attention cues (Section 3.2). This blending preserves the expressive diversity of the model while reinforcing consistency with the reference panel. Together, LPA and RAVM enable coherent character synthesis and expressive scene composition without modifying the model

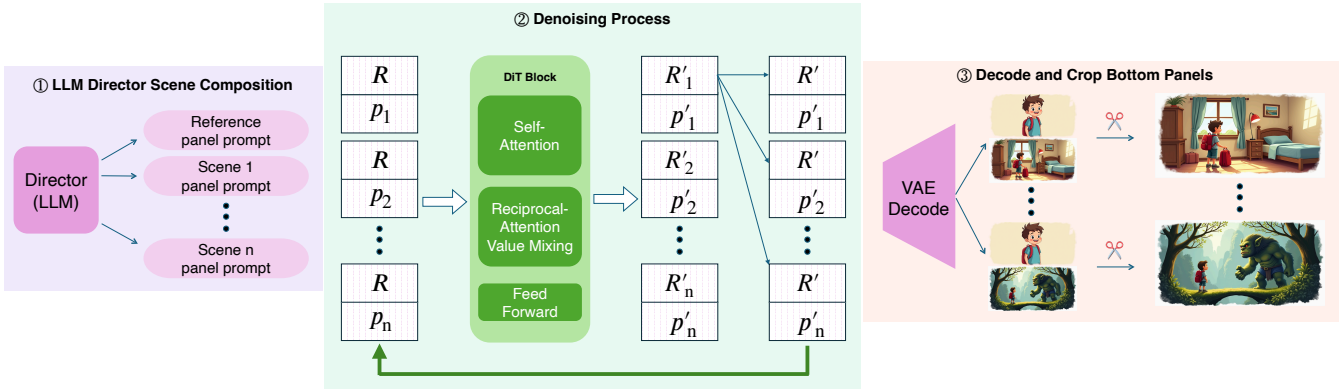

Figure 3: *Overview of our training-free storyboard generation pipeline. Given a natural language narrative (e.g., "Once upon a time, a boy set off on an adventure..."), our method proceeds in three stages: (1) An LLM-based "Director" decomposes the story into a shared reference panel prompt and a sequence of scene-level prompts; (2) A batch of n two-panel images is generated, with the top half of each image conditioned on the (same) reference prompt and the bottom half on one of the scene prompts. During denoising, we apply Latent Panel Anchoring (LPA): after each transformer block, the latent representations $[R, p_i]$ evolve to $[R'_i, p'_i]$, and the top half of each latent is replaced with the version from the first batch element, denoted $R' = R'_1$, to ensure a synchronized anchor across scenes. Inside each transformer block, we also apply Reciprocal Attention Value Mixing (RAVM) following the self-attention computation. (3) The final denoised latents are decoded into two-panel images and cropped to retain only the bottom sub-panels as the final storyboard.*

or training procedure. Implementation details and prompt schemas appear in Appendix B; pseudocode for LPA and RAVM is in Appendix C.

3.1. Latent Panel Anchoring

Our method generates a sequence of storyboard panels from a narrative text input, with consistent character identity and layout diversity across scenes. We begin by using an LLM [OAA*24] to decompose the input into a single *reference* prompt describing all recurring characters/objects and a sequence of *n scene* prompts, one per panel. These prompts are paired as described below and fed to a pre-trained text-to-image diffusion model.

For each storyboard panel we render a *two-part image* (top/bottom): the *top* half depicts the reference, and the *bottom* half depicts the scene. Concretely, we structure the text as: "A storyboard of [reference prompt] (top) and [scene prompt] (bottom)," which explicitly instructs the model what to draw in each half. Because the output image is divided into two halves, the internal latent is likewise partitioned: the top latent R evolves into the reference sub-panel, and the bottom latent p_i into the target scene (see Figure 3).

We construct a batch of n such two-part latents (one per scene) so all targets share the same reference. During denoising, each DiT block applies self-attention *within each latent* over all tokens of that latent, so the per-latent reference halves R'_i naturally drift apart as they are updated independently. To keep a single, synchronized depiction of the reference across the batch, we therefore *after each transformer block* overwrite all top halves R'_i with the top half from the first item (R'_1). After generation, we crop away the reference halves and retain the n bottom sub-panels as the final storyboard.

As illustrated in Figure 5, transformer-based diffusion models exhibit structured attention behavior: tokens corresponding to the same object–such as a character's hair, clothing, or limbs–tend to form tight clusters in key-space. These internal "cliques" facilitate soft feature sharing between semantically aligned tokens, even when they are spatially distant in the image. This behavior was previously demonstrated in the older UNet-based models [TGBD23; GBBD23; AGC*24; TKG*24b].

This consistency mechanism enables the model to propagate texture and style within and across panels. Latent Panel Anchoring leverages this emergent structure by placing a reference depiction of the characters in every latent grid, allowing attention layers to align and blend visual features between the top and bottom sub-panels.

While prompt-guided anchoring provides a strong bias toward consistency, it can be insufficient under large pose variation or complex layouts. We therefore add a complementary token-level mechanism.

3.2. Reciprocal Attention Value Mixing (RAVM)

While Latent Panel Anchoring encourages high-level visual consistency across storyboard panels, it may fail to preserve fine-grained identity, particularly when characters appear in different poses or spatial arrangements. *Reciprocal Attention Value Mixing (RAVM)* addresses this by reinforcing cross-panel correspondences between semantically aligned tokens through soft feature blending.

In attention-based diffusion models, each token's representation is updated using three components: keys, queries, and values. Prior works [TGCA23; KZZ*23] have established that keys and queries influence spatial layout and attention weighting, while values encode fine-grained visual detail such as texture, color, and appearance. RAVM acts *only* on value vectors, leaving keys/queries, and

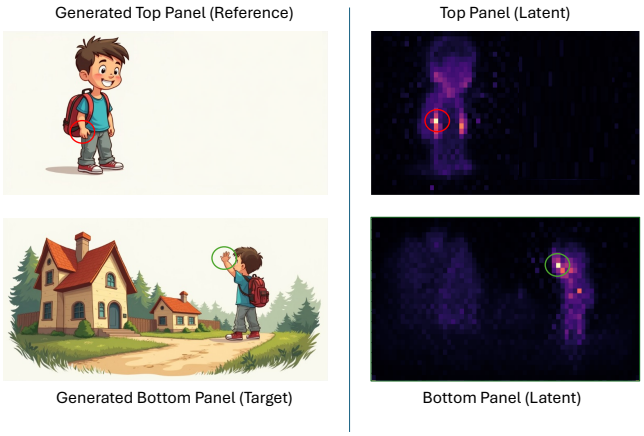

Figure 4: *Visualization of Reciprocal Attention Value Mixing (RAVM) in action.* ***Left:*** *A generated 2-panel output from our method, with the top panel serving as the shared reference. The red and green circles mark semantically corresponding character features (the hand) in the reference and target panels, respectively.* ***Right:*** *Heatmaps showing reciprocal attention scores at denoising step 12 of 28. Top-right: for each token in the top panel, we compute its reciprocal attention with the green-circled token in the bottom panel. Bottom-right: the reverse: each token in the bottom panel is scored based on reciprocal attention with the red-circled token in the top panel. In both cases, the hand token in the opposite panel receives the strongest reciprocal attention, validating that RAVM successfully identifies s e mantically a l igned t o ken p a irs f o r value mixing. This reinforces visual consistency without altering spatial composition. For an in depth explanation see Appendix F.1.*

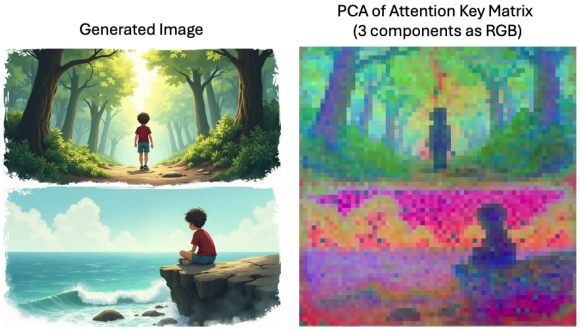

Figure 5: *Semantic key clustering within a two-panel Flux-generated storyboard (left). We extract the key vectors for each text token from a mid-layer transformer block at diffusion step 12/28, project them to three principal components, and visualize the resulting 3D embedding as RGB values (right). Tokens associated with the character (e.g., hair, face, clothing) form tight, panel-consistent clusters in key space, suggesting that the model learns a shared representation for character attributes across panels. This enables self-attention to propagate consistent texture and style cues between panels. In contrast, background-related tokens exhibit a broader spread and weaker clustering, indicating more heterogeneous key representations across the scene.*

thus attention routing and layout, unchanged. As a result, it primarily injects an appearance-level residual that reinforces consistency while preserving the base model's compositional choices, thereby preserving layout diversity.

RAVM avoids batch-level attention sharing (which may not transfer cleanly to modern DiT backbones [FHL*25]) and is applied *before* positional encoding (RoPE) to match same-character tokens without a location prior; post-RoPE mixing can degrade correspondences in some cases.

To decide *which* value vectors to mix, we identify pairs of tokens that attend strongly to each other across stacked panels. These reciprocal relationships frequently emerge between semantically aligned regions, such as a character's face or clothing, and are a key reason Latent Panel Anchoring works effectively (see Figure 5). We make this structure explicit by interpreting the two-panel latent as a *directed bipartite attention graph*: one set of nodes corresponds to tokens in the reference sub-panel, the other to tokens in the target sub-panel, and edge weights are given by attention values. We define a r eciprocal a ttention s core f or e ach c ross-panel t oken pair as the minimum of the attention in both directions, and selectively blend value vectors for those with the highest mutual connectivity.

This approach allows RAVM to softly propagate texture and style between corresponding regions, reinforcing visual consistency without overriding spatial variation or requiring explicit supervision.

Formally, let $x \in \mathbb{R}^{2P\times d}$ denote the concatenated tokens of the reference and target sub-panels. The model computes:

$$Q = xW_Q, \quad K = xW_K, \quad A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{2P\times 2P}, \quad (1)$$

where $A[i,j]$ is the attention from token i to token j. We extract the cross-panel blocks:

$$A_{tb} = A[1{:}P,\ P{:}2P], \qquad A_{bt} = A[P{:}2P,\ 1{:}P], \qquad (2)$$

corresponding to top-to-bottom and bottom-to-top attention.

We define the *reciprocal attention score* between a top token u and bottom token v as:

$$\text{RA}(u,v) := \min(A_{tb}[u,v], A_{bt}[v,u]), \qquad (3)$$

yielding a symmetric matrix $M \in \mathbb{R}^{P\times P}$ of bidirectional scores, which during inference can be efficiently computed for all tokens by:

$$M := \min(A_{bt}, A_{tb}^T) \qquad \text{(elementwise)} \qquad (4)$$

We maintain an exponential moving average $\bar{M}$ of M across transformer layers and diffusion steps.

To extract high-confidence correspondences, we analyze the weights on the edges crossing the bipartite graph cut. We apply Otsu's thresholding method [Ots79] to the reciprocal attention matrix $\bar{M}$ and use morphological filtering to clean the resulting binary

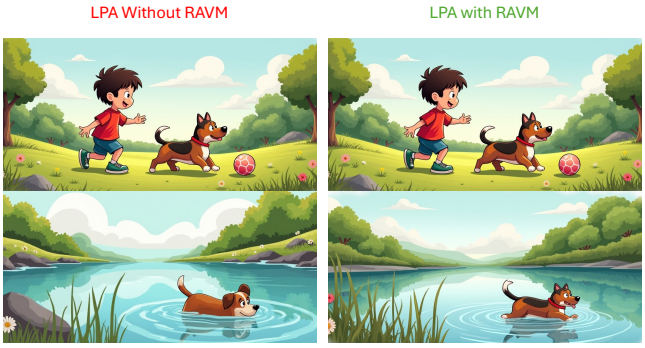

Figure 6: ***Effect of RAVM on identity consistency.*** *Each column shows a two-part storyboard: top (reference) "a boy playing with his dog"; bottom (scene) "the dog swimming in a lake."* Left *(LPA without RAVM): weaker appearance consistency across panels.* Right *(LPA with RAVM): value-space blending guided by reciprocal attention refines appearance.* ***Character selection.*** *Because the scene prompt mentions only the dog, the boy is naturally omitted from the bottom panel, and no identifier tokens or masks are required. LPA specifies the bottom via natural language, and RAVM follows the model's internal attention rather than imposing new routing.*

mask. For each selected bottom token v, we identify the top-panel token u^* with the highest reciprocal score:

$$u^* = \arg\max_u \bar{M}[u, v], \tag{5}$$

and apply a soft value update:

$$V'_v = \lambda V_v + (1-\lambda) V_{u^*}, \tag{6}$$

where V_i is the value vector of token i, and λ is a mixing weight. Since keys and queries remain unchanged, the spatial layout and attention dynamics of the scene are preserved.

By reinforcing only the strongest reciprocal connections across the attention graph, RAVM enhances character consistency without suppressing scene diversity or altering the model's generative flexibility.

4. Experiments

We evaluate our method both qualitatively and quantitatively, focusing on three core dimensions: prompt alignment, character consistency, and scene diversity. To support this, we introduce the **Rich Storyboard Benchmark**, designed to test narrative and compositional expressiveness beyond the scope of existing identity-focused datasets. We also evaluate on the DS-500 benchmark [HYT*24] to demonstrate generalizability.

Section 4.1 outlines the methods compared; Section 4.2 details our benchmark and metrics. Results are reported in Sections 4.3 and 4.4, with human preference scores in Section 4.5.

4.1. Baselines and Comparison Setup

We compare against story-centric models: **StoryDiffusion** [ZZC*24], which introduces Consistent Self-Attention; **ConsiStory** [TKG*24b], an extended-attention method that shares attention across frames; **IC-LoRA** [HWW*24], evaluated

Figure 7: *A four-panel storyboard featuring Blackpaw, a shimmering fox of the ancient celestial forest. Each scene is grounded in a specific scene from a longer story (full text in Appendix J). Our method preserves character consistency while supporting expressive spatial framing and richly atmospheric environments. Even as Blackpaw varies in pose, size, and placement across panels, the evolving backgrounds remain narratively grounded and visually coherent. Key visual moments are drawn from the following excerpts: "... With a flick of his glowing tail, he bounded across a fallen tree stretched precariously over a mist-shrouded ravine that gleamed faintly ... Perched atop a broken archway of ancient stone, vines and silver moss hanging around him... ... From the edge of a luminous lake mirroring the heavens perfectly, he watched a meteor shower ignite the sky... ... Curling beside a pulsing crystal monolith, he dreamed..."*

in storyboard and movie-shot finetuned variants; and **StoryGen** [LWZ*24], an autoregressive generator driven by scene prompts. We also include **OminiControl** [TLY*24], which leverages a trained image encoder and a reference image to guide layout and style, outperforming prior encoder-based methods. In addition, we evaluate **Flux.1 Kontext** [LBB*25], which conditions on concatenated prompts and/or reference images to propagate identity across panels, providing a strong consistency-oriented baseline.

For ablations, we evaluate the Flux base model (Flux.1-dev [Bla24]) (no consistency mechanisms), and a version with Latent Panel Anchoring (LPA) only, isolating the contribution of Reciprocal Attention Value Mixing (RAVM). We also experiment with varying the value mixing coefficient λ to assess its effect on consistency and expressiveness, and a top k neighborhood variant of RAVM.

4.2. Benchmark and Evaluation Metrics

Rich Storyboard Benchmark. Existing benchmarks focus primarily on identity preservation and do not reflect the compositional or cinematic demands of visual storytelling. To address this gap, we introduce the Rich Storyboard Benchmark, a set of 100 open-domain story prompts, each decomposed into seven richly detailed scene-level descriptions. The benchmark emphasizes dynamic layout, spatial diversity, and character-scene interaction, all critical for assessing visual narrative quality beyond identity fidelity. We provide detailed information about the benchmark's construction and composition in Appendix A.

Metrics. We evaluate prompt alignment using VQAScore [LPL*24], character consistency using DreamSim [FTS*23], and scene diversity using our novel **Scene Diversity** metric, which measures how dynamically the subject is presented across panels (changes in framing, position, scale, and pose). Given a storyboard of n panels and a text description identifying the subject, we locate the subject in each image using Grounding DINO [LZR*24]. For each panel, we extract a bounding box around the subject, normalize it by the image dimensions, and compute the per-coordinate standard deviation of these normalized bounding boxes across the n panels. We then average these standard deviations to obtain a bounding box std score, reflecting variation in subject placement and scale. Each story's bounding box std score is min-max normalized across all stories in the benchmark to yield s_{bbox}. For stories with human characters, we additionally compute 17 pose keypoints per panel using ViTPose [XZZT22]. We calculate the per-keypoint variance across all panels and average them to obtain a pose variance score, which is then min-max normalized across the benchmark to yield s_{pose}. Our final score is:

$$\text{Scene Diversity} = \begin{cases} \frac{1}{2}(s_{\text{bbox}} + s_{\text{pose}}), & \text{if human} \\ s_{\text{bbox}}, & \text{otherwise} \end{cases}$$

This metric enables us to evaluate a model's ability to vary subject framing and presentation across narrative beats, a core requirement for expressive visual storytelling. We report additional metrics in Appendix I.

As an attention diagnostic, we also measure average cross-panel attention mass from bottom-panel tokens to the reference panel, separately for foreground vs. background regions (Appendix G).

4.3. Qualitative Evaluation

Full Storyboard Comparison. We present 4-panel storyboard sequences for two representative stories rendered by each method (Figures 7 and 11). Our method achieves a stronger balance across prompt alignment, character consistency, and scene diversity. It supports varied framing and character positioning while maintaining coherent, visually rich environments. Baselines tend to overfit one aspect: ConsiStory and StoryDiffusion favor centered subjects; IC-LoRA repeats compositional templates; and OminiControl often omits off-center characters. Our method accommodates these challenges, yielding coherent, expressive storyboards.

4.4. Quantitative Evaluation

We evaluate on both the Rich Storyboard Benchmark and DS-500 [HYT*24]. Metrics are computed per storyboard and averaged.

Figure 8 shows that our method dominates the Pareto front in prompt alignment and character consistency. We visualize metrics pairwise to better expose tradeoffs: for instance, models with high character consistency often achieve it by sacrificing prompt alignment or layout flexibility. The Flux baseline exemplifies this pattern: it attains strong consistency scores by rendering nearly identical characters across panels, but lacks sensitivity to prompt-specific content, resulting in lower alignment and limited scene variation.

DS-500 Evaluation. To assess generalization beyond our benchmark, we also evaluate on DS-500 [HTC*25], a storyboard dataset with shorter prompts and minimal scene evolution. While not designed to test layout or narrative expressivity, DS-500 remains a useful baseline for identity coherence. Our method performs competitively on this benchmark, despite its focus on richer visual storytelling. Full results and comparisons are provided in Appendix D.

Ablation Study. All figures include two ablations of our method: a vanilla Flux baseline (no consistency mechanisms), and Flux with LPA only. Our full method outperforms both, highlighting the complementary roles of LPA and RAVM. While LPA alone improves layout coherence, RAVM significantly boosts character consistency without harming layout diversity.

We further analyze the effect of the mixing parameter λ in the RAVM update. Increasing λ leads to a clear improvement in character consistency, as stronger blending amplifies the influence of semantically aligned reference tokens. Interestingly, we also observe gains in prompt alignment and scene diversity. While RAVM only modifies the value vectors, which capture texture and fine appearance details, we hypothesize that stabilizing the character's appearance helps the model more clearly separate and render background elements across panels, resulting in richer scenes and improved prompt fidelity.

We also study a top k neighborhood variant of RAVM that uniformly averages values from the k highest reciprocal matches (no dynamic weights). Increasing $k>1$ slightly lowers consistency and yields no measurable gains in alignment/diversity (see Appendix E for more details).

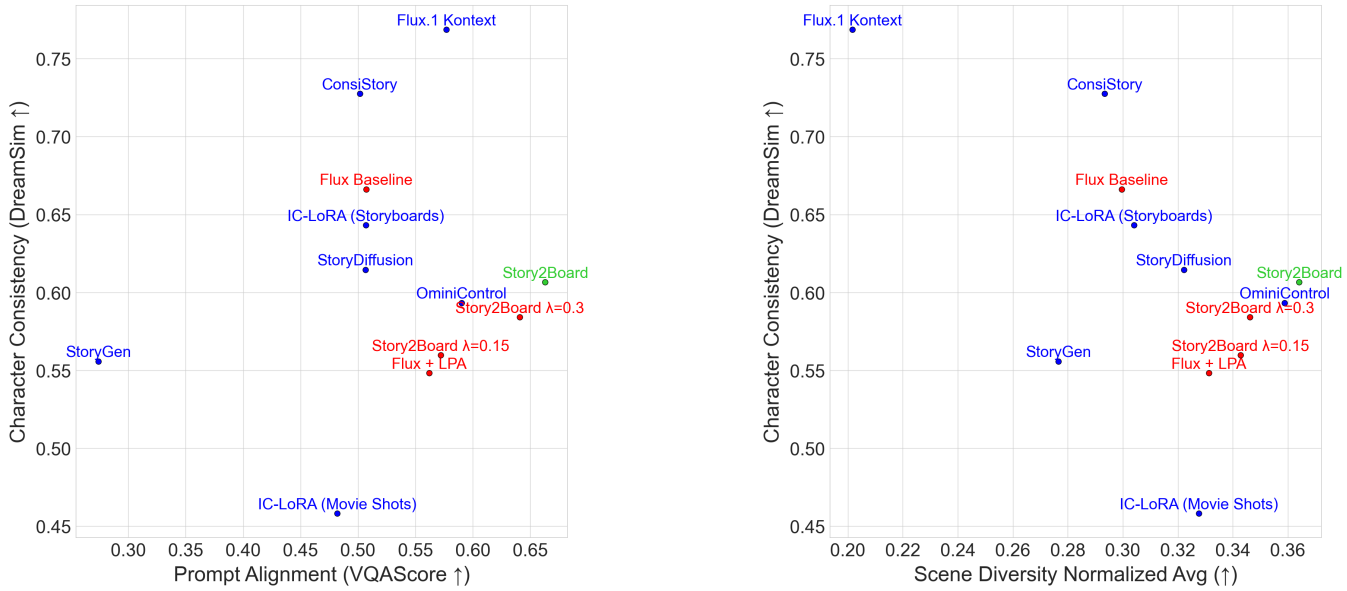

Figure 8: ***Left: Character Consistency vs. Prompt Alignment.*** *Story2Board achieves the best tradeoff, outperforming all baselines and ablations. Prompt alignment (x-axis) is measured via VQAScore and character consistency (y-axis) via DreamSim. The Flux baseline exhibits unusually high consistency due to its collapsed behavior–rendering similar characters across panels with minimal pose or appearance variation–yet struggles with prompt grounding.* ***Right: Scene Diversity vs. Character Consistency.*** *Our method maintains high identity fidelity while enabling significantly more layout variation than competing methods. Scene Diversity (x-axis) is our proposed metric (details in supplementary), while character consistency (y-axis) is again measured via DreamSim. Note that IC-LoRA baselines (Movie Shots and Storyboards) operate only on 4-panel sequences and are not applicable to longer formats.*

4.5. User Study

To supplement our quantitative evaluation, we conducted a large-scale user study via the Amazon Mechanical Turk (AMT) platform [Ama25], using all 100 stories from our Rich Storyboard Benchmark. For each story, we generated 4-panel storyboards using our method and each of the competing baselines. Each worker task consisted of a pairwise comparison between two storyboards (one from our method and one from a baseline), with each comparison focused on one of five criteria: overall preference, prompt alignment, character consistency, background richness, and scene diversity. In total, 700 such tasks were created, and each was completed by three independent workers. Figure 9 summarizes the results, showing the preference rate of our method over each baseline for the five criteria. More details about the study can be found in Appendix H.

Our method was the most preferred overall, winning the majority of pairwise comparisons in the "Overall Preference" category. This suggests that when users evaluated storyboards holistically, they consistently favored our approach over all baselines.

These outcomes reflect an inherent trade-off between prompt alignment/scene diversity and character consistency: a good storyboard method must strike the right balance. OminiControl wins in prompt alignment, background richness, and scene diversity, but loses in character consistency; it pushes too far toward compositional control, which in turn yields a lower overall preference. Conversely, Flux.1 Kontext, StoryDiffusion, IC-LoRA (Storyboards) and Flux.1-dev score higher on consistency yet lose scene diversity, often exhibiting *consistency-by-repetition* (i.e., the character remains nearly unchanged in pose/scale/placement across panels.). Story2Board achieves the best overall preference precisely because it balances these competing axes: even when a baseline outperforms us on a single criterion, the combined viewing experience is stronger with our method.

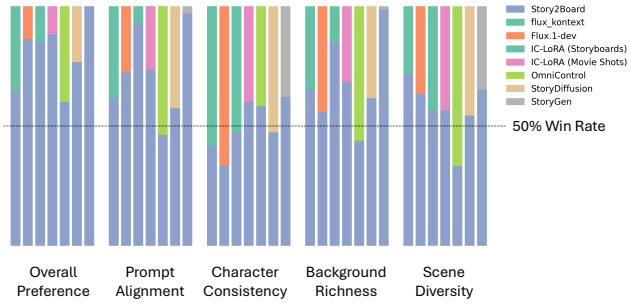

Figure 9: ***User Study Results.*** *Participants compared Story2Board to competing systems across five evaluation dimensions. Our method is preferred overall, and achieves strong performance across all categories. While some baselines edge ahead in isolated metrics, Story2Board strikes the best balance between character consistency, visual richness, and narrative alignment, a key strength for storyboard generation.*

Figure 10: ***Attention Entanglement in Flux.*** *Left: A two-panel storyboard generated by Flux without our method. Attention entanglement causes the fairy to erroneously inherit the raccoon's tail in the top panel, while in the bottom panel the raccoon adopts the fairy's wings. Right: With our Mutual Attention (MA) mechanism, these misattributions persist, but their visual appearance becomes consistent across panels. MA also improves the consistency of other visual elements, such as the raccoon's tail and the lantern, demonstrating the broader stabilizing influence of token-level value mixing. When entangled representations are already present in the base model, our method propagates rather than corrects them.*

4.6. Limitations

Our method leverages the base model's attention dynamics, and therefore inherits its failure modes. One well-documented issue is *attention entanglement*, including phenomena such as incorrect attribute binding, object fusion, and semantic misassignment, where separate concepts or entities interfere with one another during generation [DPAC24]. Since our approach reinforces token-level correspondences based on reciprocal attention, it may propagate these entanglements if they persist during denoising; it cannot reliably undo them once they form. An example is shown in Figure 10. Second, RAVM "rides on" the the model's reciprocal cross-panel links: if no strong reciprocal support emerges for a subject, RAVM has little signal to mix. In practice, this limitation is proportional to the number or characters (e.g., more than two), in which case additional prompt engineering may be required.

5. Summary and Discussion

Advantages over prior approaches. Beyond consistency and diversity, our pipeline offers two practical advantages. First, it supports *subset selection*: when the shared reference depicts multiple entities, a scene can mention only a subset and the model will render just those entities (Figure 6). This emerges because LPA specifies the bottom panel *purely via natural language*, leveraging the base model's flexibility to compose the scene, while RAVM *does not impose new attention routing*: unlike extended/shared-attention approaches that couple panels by sharing attention maps or Q/K interactions, we preserve the model's routing and blend only appearance information in the attention *values* where reciprocal support is high, thus "not getting in the way." Second, unlike methods that rely on explicit subject annotations and attention masks (e.g., DreamStory [HYT*24]), our approach operates directly on token-level features without external supervision, remaining training-free and open-domain.

We introduced Story2Board, a training-free framework for generating visually consistent and compositionally rich storyboards from text. By leveraging reciprocal attention patterns between tokens across panels, our method reinforces character identity while preserving layout diversity. Through extensive experiments on our proposed benchmark and a standard existing dataset, we demonstrate that Story2Board enables more dynamic, expressive visual storytelling than prior approaches.

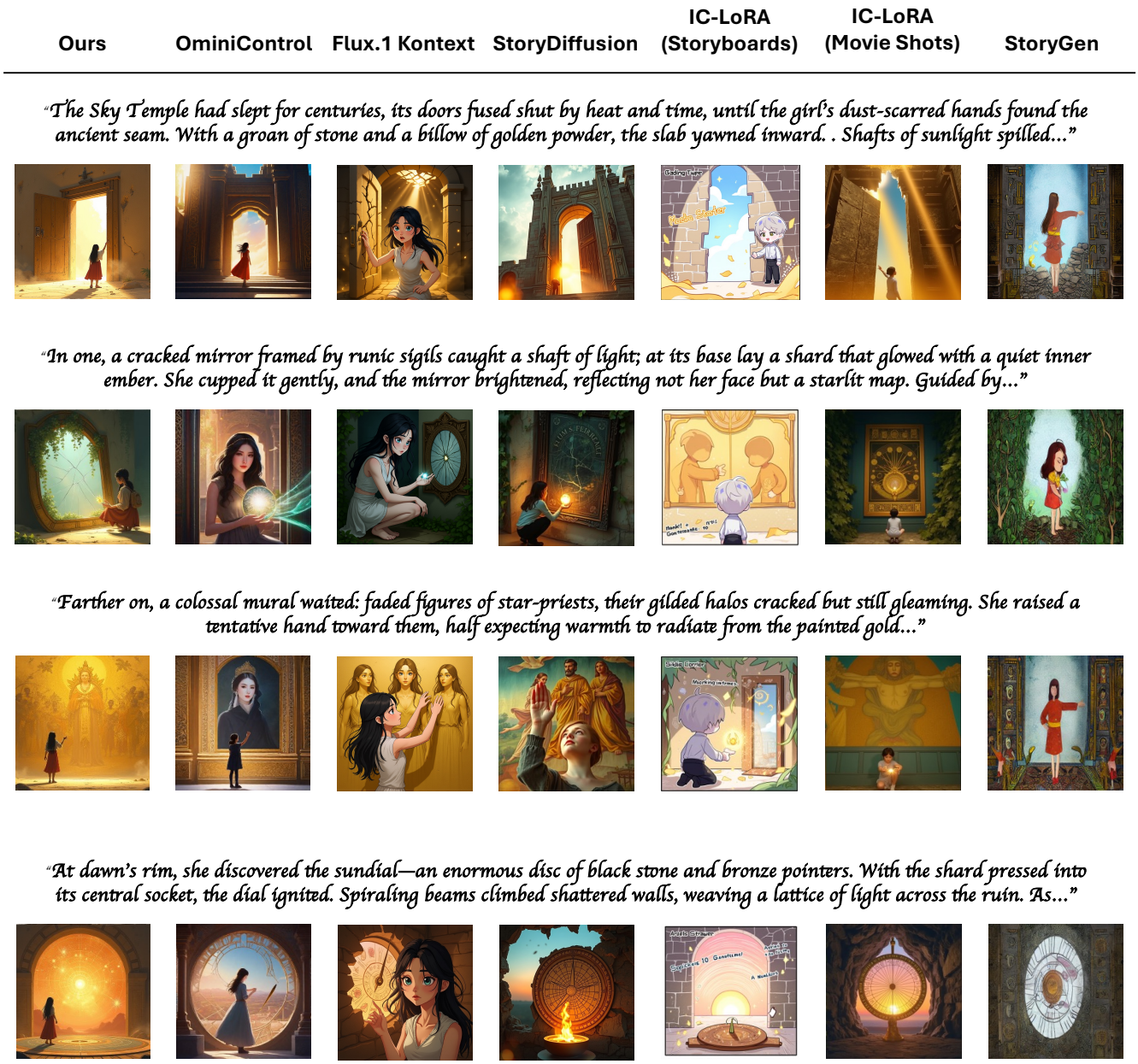

Figure 11: ***Qualitative comparison of multi-panel storyboards.*** *Our method (*STORY2BOARD*, left column) achieves a three-way balance:* scene diversity *(varying viewpoints, scale, and richly grounded backgrounds),* character consistency *(stable appearance and silhouette), and tight* prompt alignment*. Baseline systems each miss at least one of these axes:* STORYDIFFUSION *varies layouts but allows the heroine's features to drift;* OMINICONTROL *creates atmospheric backdrops yet occasionally omits the protagonist;* IC-LORA STORYBOARDS *fixes the camera and produces stylised cartoon frames, limiting narrative variety;* IC-LORA MOVIE SHOTS *shows wider layouts but often mismatches prompt details;* STORYGEN *produces stylized frames, but struggles with narrative continuity and compositional coherence across panels. By maintaining identity while continuously re-contextualising the scene,* STORY2BOARD *delivers the most faithful and visually engaging storyboard among current approaches.*

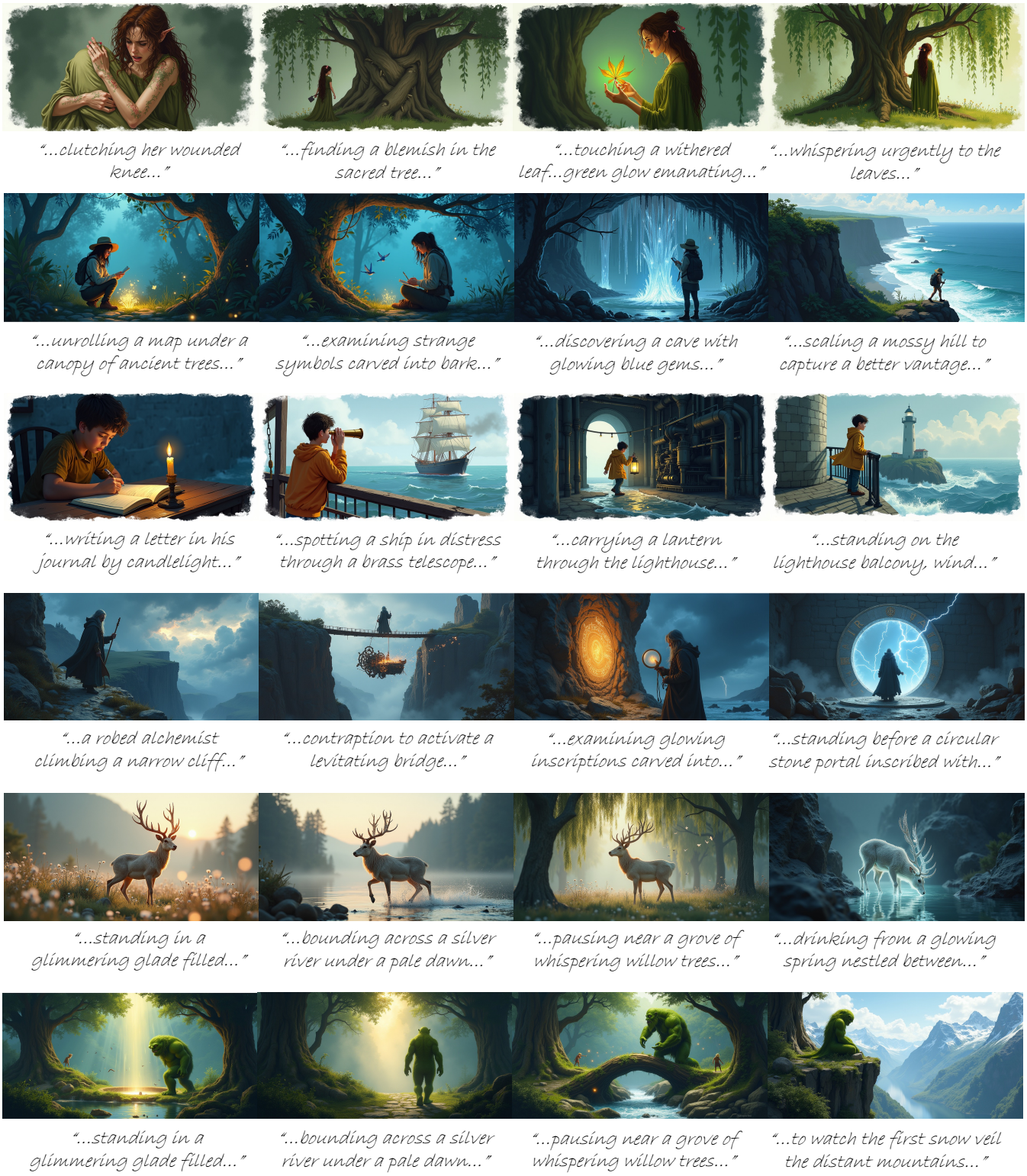

Figure 12: *Additional storyboards generated by our method.*

References

[AGC*24] AVRAHAMI, OMRI, GAL, RINON, CHECHIK, GAL, et al. "DiffUHaul: A Training-Free Method for Object Dragging in Images". *SIGGRAPH Asia 2024 Conference Papers*. SA '24. Association for Computing Machinery, 2024. ISBN: 9798400711312. DOI: 10.1145/3680528.3687590. URL: https://doi.org/10.1145/3680528.3687590 4.

[AHV*24] AVRAHAMI, OMRI, HERTZ, AMIR, VINKER, YAEL, et al. "The Chosen One: Consistent Characters in Text-to-Image Diffusion Models". *ACM SIGGRAPH 2024 Conference Papers*. SIGGRAPH '24. Denver, CO, USA: Association for Computing Machinery, 2024. ISBN: 9798400705250. DOI: 10.1145/3641519.3657430. URL: https://doi.org/10.1145/3641519.3657430 3.

[Ama25] AMAZON MECHANICAL TURK. *Amazon Mechanical Turk*. https://www.mturk.com/. Accessed: 2025-05-20. 2025 8, 20.

[Ani14] ANIMATOR ISLAND. *Composition: What is Breathing Room?* https://www.animatorisland.com/composition-what-is-breathing-room/. Accessed: 2025-05-12. 2014 2.

[Bla24] BLACK FOREST LABS. *FLUX*. https://github.com/black-forest-labs/flux. 2024 2, 3, 7, 17.

[Blo20] BLOCK, BRUCE. *The Visual Story: Creating the Visual Structure of Film, TV, and Digital Media*. 3rd. Focal Press, 2020 2.

[DPAC24] DAHARY, OMER, PATASHNIK, OR, ABERMAN, KFIR, and COHEN-OR, DANIEL. "Be yourself: Bounded attention for multi-subject text-to-image generation". *European Conference on Computer Vision*. Springer. 2024, 432–448 9.

[EKB*24] ESSER, PATRICK, KULAL, SUMITH, BLATTMANN, A., et al. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". *ArXiv* abs/2403.03206 (2024) 3.

[FHL*25] FENG, HAORAN, HUANG, ZEHUAN, LI, LIN, et al. "Personalize anything for free with diffusion transformer". *arXiv preprint arXiv:2503.12590* (2025) 3, 5.

[Fil25] FILMMAKERS ACADEMY. *Negative Space: Film Composition Guide*. https://www.filmmakersacademy.com/blog-negative-space-film/. Accessed: 2025-05-12. 2025 2.

[FTS*23] FU, STEPHANIE, TAMIR, NETANEL Y., SUNDARAM, SHOBHITA, et al. "DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data". *ArXiv* abs/2306.09344 (2023) 7, 18, 21.

[GBBD23] GEYER, MICHAL, BAR-TAL, OMER, BAGON, SHAI, and DEKEL, TALI. "Tokenflow: Consistent diffusion features for consistent video editing". *arXiv preprint arXiv:2307.10373* (2023) 4.

[HJA20] HO, JONATHAN, JAIN, AJAY, and ABBEEL, PIETER. "Denoising Diffusion Probabilistic Models". *Proc. NeurIPS*. 2020 2, 3.

[HTC*25] HE, JUNJIE, TUO, YUXIANG, CHEN, BINGHUI, et al. "AnyStory: Towards Unified Single and Multiple Subject Personalization in Text-to-Image Generation". *arXiv preprint arXiv:2501.09503* (2025) 2, 3, 7.

[HWW*24] HUANG, LIANGHUA, WANG, WEI, WU, ZHIGANG, et al. "In-Context LoRA for Diffusion Transformers". *ArXiv* abs/2410.23775 (2024) 3, 6.

[HYT*24] HE, HUIGUO, YANG, HUAN, TUO, ZIXI, et al. "Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion". *arXiv preprint arXiv:2407.12899* (2024) 2, 3, 6, 7, 9, 18.

[KMR*23] KIRILLOV, ALEXANDER, MINTUN, ERIC, RAVI, NIKHILA, et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV] 3.

[KZZ*23] KUMARI, NUPUR, ZHANG, BINGLIANG, ZHANG, RICHARD, et al. "Multi-concept customization of text-to-image diffusion". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 1931–1941 4.

[LBB*25] LABS, BLACK FOREST, BATIFOL, STEPHEN, BLATTMANN, ANDREAS, et al. *FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space*. 2025. arXiv: 2506.15742 [cs.GR]. URL: https://arxiv.org/abs/2506.15742 7.

[LPL*24] LIN, ZHIQIU, PATHAK, DEEPAK, LI, BAIQI, et al. "Evaluating Text-to-Visual Generation with Image-to-Text Generation". *European Conference on Computer Vision*. 2024 7, 21.

[LWZ*24] LIU, CHANG, WU, HAONING, ZHONG, YUJIE, et al. *Intelligent Grimm – Open-ended Visual Storytelling via Latent Diffusion Models*. 2024. arXiv: 2306.00973 [cs.CV]. URL: https://arxiv.org/abs/2306.00973 2, 3, 7.

[LZL*23] LIU, SHAOTENG, ZHANG, YUECHEN, LI, WENBO, et al. "Video-p2p: Video editing with cross-attention control". *arXiv preprint arXiv:2303.04761* (2023) 2.

[LZR*24] LIU, SHILONG, ZENG, ZHAOYANG, REN, TIANHE, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection". *European conference on computer vision*. Springer. 2024, 38–55 7, 21.

[OAA*24] OPENAI, ACHIAM, JOSH, ADLER, STEVEN, et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774 4, 14.

[Ots79] OTSU, NOBUYUKI. "A Threshold Selection Method from Gray-Level Histograms". *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), 62–66. DOI: 10.1109/TSMC.1979.4310076 5.

[PEL*23] PODELL, DUSTIN, ENGLISH, ZION, LACEY, KYLE, et al. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". *ArXiv* abs/2307.01952 (2023) 2, 3.

[RBL*21] ROMBACH, ROBIN, BLATTMANN, A., LORENZ, DOMINIK, et al. "High-Resolution Image Synthesis with Latent Diffusion Models". *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685 2, 3.

[RDN*22] RAMESH, ADITYA, DHARIWAL, PRAFULLA, NICHOL, ALEX, et al. "Hierarchical text-conditional image generation with CLIP latents". *arXiv preprint arXiv:2204.06125* (2022) 2, 3.

[RKH*21] RADFORD, ALEC, KIM, JONG WOOK, HALLACY, CHRIS, et al. "Learning transferable visual models from natural language supervision". *International conference on machine learning*. PmLR. 2021, 8748–8763 18, 21.

[SCS*22] SAHARIA, CHITWAN, CHAN, WILLIAM, SAXENA, SAURABH, et al. "Photorealistic text-to-image diffusion models with deep language understanding". *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494 2, 3.

[Sta24] STABILITY AI. *Stable Diffusion 3: Next-Generation Text-to-Image Generation*. urlhttps://stability.ai/news/stable-diffusion-3. 2024 2.

[TGBD23] TUMANYAN, NAREK, GEYER, MICHAL, BAGON, SHAI, and DEKEL, TALI. "Plug-and-play diffusion features for text-driven image-to-image translation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 1921–1930 4.

[TGCA23] TEWEL, YOAD, GAL, RINON, CHECHIK, GAL, and ATZMON, YUVAL. "Key-Locked Rank One Editing for Text-to-Image Personalization". *ACM SIGGRAPH 2023 Conference Proceedings*. SIGGRAPH '23. Los Angeles, CA, USA, 2023 4.

[TKG*24a] TEWEL, YOAD, KADURI, OMRI, GAL, RINON, et al. "Training-Free Consistent Text-to-Image Generation". *ArXiv* abs/2402.03286 (2024) 3.

[TKG*24b] TEWEL, YOAD, KADURI, OMRI, GAL, RINON, et al. "Training-free consistent text-to-image generation". *ACM Transactions on Graphics (TOG)* 43.4 (2024), 1–18 2, 4, 6.

[TLY*24] TAN, ZHENXIONG, LIU, SONGHUA, YANG, XINGYI, et al. "OminiControl: Minimal and Universal Control for Diffusion Transformer". *arXiv preprint arXiv:2411.15098* (2024) 2, 3, 7.

[VSP*17] VASWANI, ASHISH, SHAZEER, NOAM, PARMAR, NIKI, et al. "Attention is all you need". *Advances in neural information processing systems* 30 (2017) 3.

[WZJ*23] WEI, YUXIANG, ZHANG, YABO, JI, ZHILONG, et al. "ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation". *ArXiv* abs/2302.13848 (2023) 2.

[XZZT22] XU, YUFEI, ZHANG, JING, ZHANG, QIMING, and TAO, DACHENG. "Vitpose: Simple vision transformer baselines for human pose estimation". *Advances in neural information processing systems* 35 (2022), 38571–38584 7.

[YGL*24] YANG, SHUAI, GE, YUYING, LI, YANG, et al. "Seed-story: Multimodal long story generation with large language model". *arXiv preprint arXiv:2407.08683* (2024) 2.

[YZL*23] YE, HU, ZHANG, JUN, LIU, SIBO, et al. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models". *arXiv* abs/2308.06721 (2023) 2, 3.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A., et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 586–595 21.

[ZMC*23] ZHANG, KAI, MO, LINGBO, CHEN, WENHU, et al. "MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing". *Advances in Neural Information Processing Systems*. 2023 2.

[ZZC*24] ZHOU, YUPENG, ZHOU, DAQUAN, CHENG, MING-MING, et al. "StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation". *ArXiv* abs/2405.01434 (2024) 2, 3, 6.

Story2Board: Appendix

A. Rich Storyboard Benchmark

To evaluate expressive visual storytelling, we introduce the **Rich Storyboard Benchmark**, a collection of 100 richly detailed and narratively structured short stories. Prompts were generated with GPT–4o [OAA*24], using category–specific instructions that emphasize narrative progression, visual diversity, and character continuity.

Benchmark structure (what is included). The benchmark is partitioned into three sub–benchmarks, each probing a different kind of identity consistency:

1. **Single Character:** one persistent protagonist across all panels.
2. **Character & Strangers:** one persistent protagonist interacting with varying, non–persistent side-characters ("strangers").
3. **Character & Companion:** two persistent protagonists that appear both together and separately across panels.

All entries require evolving backgrounds (lighting, weather, architecture, etc.) across panels. We balance human, animal, and fantasy settings and vary scene types. Illustrative examples of “Characters and Strangers” appear Figure A1.

A.1. Benchmark Construction

We programmatically generated 100 entries with GPT–4o from category–specific instructions (below), then performed a light manual pass to remove low–quality, unsafe, or unethical items and to enforce our schema (evolving backgrounds, category constraints, etc.). In this appendix we document the prompts used to *construct the dataset*. (Implementation details of any LLM components used during *model inference* are described separately in §B.)

GPT–4o Prompt (used to generate *Single Character* entries)

> You are a storyboard designer helping to create cinematic, visually expressive storyboards. Your task is to:
> (1) Invent a short fantasy or human-centric story with a clear and exciting story arc.
> (2) Make sure the story contains not only a memorable character, but also richly described background elements that evolve across scenes. These might include weather, lighting, ruins, architecture, mountains, oceans, caves, temples, streets, or crowds. Do not focus only on the character.
> (3) Break down the story into seven distinct storyboard scene descriptions. Each should describe one key visual moment–something worth illustrating in a single panel. Replace names with a description of the character (e.g. instead of "Tom went to the store", say "A man went to the store").
> Return:
> - The full story
> - A list of 7 storyboard panel prompts that visually represent the story

Example Output (from GPT-4o)

> **Story:** Elias, the lone sailor in a battered wool cap and dark coat, stood calmly at the bow of his wooden ship, its prow cutting through the golden morning fog. The endless sea stretched around him, a liquid desert beneath a flickering sky. Later, Elias sat on a weathered crate beneath heavy coils of rope, gulls circling overhead, the ship creaking with age. He gazed outward, where the vastness made even the towering masts seem like toys adrift in a dream. At dusk, he leaned against the rail, lanterns swinging from the rigging above him as he stared at the glowing horizon, where the sky burned into the water. A massive arch of stone rose from the ocean ahead, so gigantic that Elias felt as small as a pebble beside it, looking up in awe. When night fell, he rested beside the ship’s wheel, moonlight painting silver paths across the undulating waves. Eventually, only a quiet harbor awaited him, ropes swaying in the wind, the ship itself nowhere to be seen–as if it had drifted into legend.
>
> **Storyboard panel prompts:**
>
> 1. A lone sailor in a wool cap and dark coat standing calmly at the bow of a wooden ship in golden morning fog
> 2. A lone sailor in a wool cap and dark coat sitting on a crate beneath coiled ropes as gulls circle overhead
> 3. A lone sailor in a wool cap and dark coat gazing at the open sea, with lanterns swaying gently from the rigging
> 4. A lone sailor in a wool cap and dark coat standing on the deck smiling as the ship approaches a massive arch of stone rising from the water
> 5. A lone sailor in a wool cap and dark coat resting beside the ship’s steering wheel, moonlight shimmering on the sea
> 6. A lone sailor in a wool cap and dark coat guiding the ship through towering jade waves as bioluminescent whales breach alongside
> 7. A lone sailor in a wool cap and dark coat playing a low whistle tune on the bow beneath a sky ablaze with meteoric shards

GPT–4o Prompt (used to generate *Character & Strangers* entries)

> You are a storyboard designer helping to create cinematic, visually expressive storyboards. Your task is to:
> (1) Invent a short human, animal, or fantasy–centric story centered on one persistent protagonist. You may use names.
> (2) Across the seven panels, have the protagonist interact with different strangers (bystanders, traders, pilgrims, guides, performers, etc.). Exactly four of the seven panels should include an interaction with a stranger (each stranger appears in at most one panel). The remaining three panels should feature the protagonist alone.
> (3) Ensure evolving, richly described backgrounds across scenes (lighting, weather, terrain, architecture, crowds), with varied poses and camera angles for the protagonist.
> (4) Provide seven distinct panel descriptions, each a single visually striking moment that implies interaction or solitary intent as specified above.

Figure A1: ***Characters and Strangers.*** *Each row shows a storyboard from the Rich Storyboard Benchmark where the main character encounters unfamiliar figures, testing the model's ability to maintain character identity while integrating diverse background elements.*

> Return: (a) the full story (one or two short paragraphs) and (b) a list of exactly 7 storyboard panel prompts.

Example Output (Character & Strangers): ***The Cartographer's Compass*** **Story.** A young cartographer set out with a leather satchel and a brass compass, chasing the rumor of a vanished road that threaded mountain, desert, and shore. In a windblown field he tested the needle against distant peaks; on a beach, an old woman showed him maps carved into driftwood, sea–grayed and true. In a cavern lit by glowing moss, he traced constellations on the ceiling, listening to the sea thrum through stone. At a desert crossroads, a cloaked traveler traded stories beneath hanging lanterns that hummed with moths. The cartographer found a half–buried stone gate and knelt to read its runes, sand whispering over the carvings. A drifting poet taught him to read the wind by petals skimming the air. At last, beneath a comet–scarred sky, a silent pilgrim's lantern mirrored the crater's glow, and the compass needle settled—pointing not to a place, but to a way.

Storyboard panel prompts.

1. A young cartographer with a leather satchel and brass compass standing in a windblown field, holding up a brass compass and studying distant peaks
2. A young cartographer with a leather satchel and brass compass sketching beside an old woman who shows him maps carved into driftwood on a beach
3. A young cartographer with a leather satchel and brass compass tracing constellations on a cave ceiling lit by glowing moss, alone with the echoing tide
4. A young cartographer with a leather satchel and brass compass trading stories with a cloaked traveler beneath hanging lanterns at a desert crossroads
5. A young cartographer with a leather satchel and brass compass crouching alone to examine ancient runes etched into a stone gate half–buried in sand
6. A young cartographer with a leather satchel and brass compass mapping gust–drawn trails of migrating petals with a drifting poet who reads the winds
7. A young cartographer with a leather satchel and brass compass marking a star–scarred crater's rim beside a silent pilgrim whose lantern mirrors the comet's glow

GPT–4o Prompt (used to generate *Character & Companion* entries)

> You are a storyboard designer helping to create cinematic, visually expressive storyboards. Your task is to:
> (1) Invent a short human, animal, or fantasy–centric story featuring two persistent protagonists (e.g., a person and their animal/robot companion). You may use names.

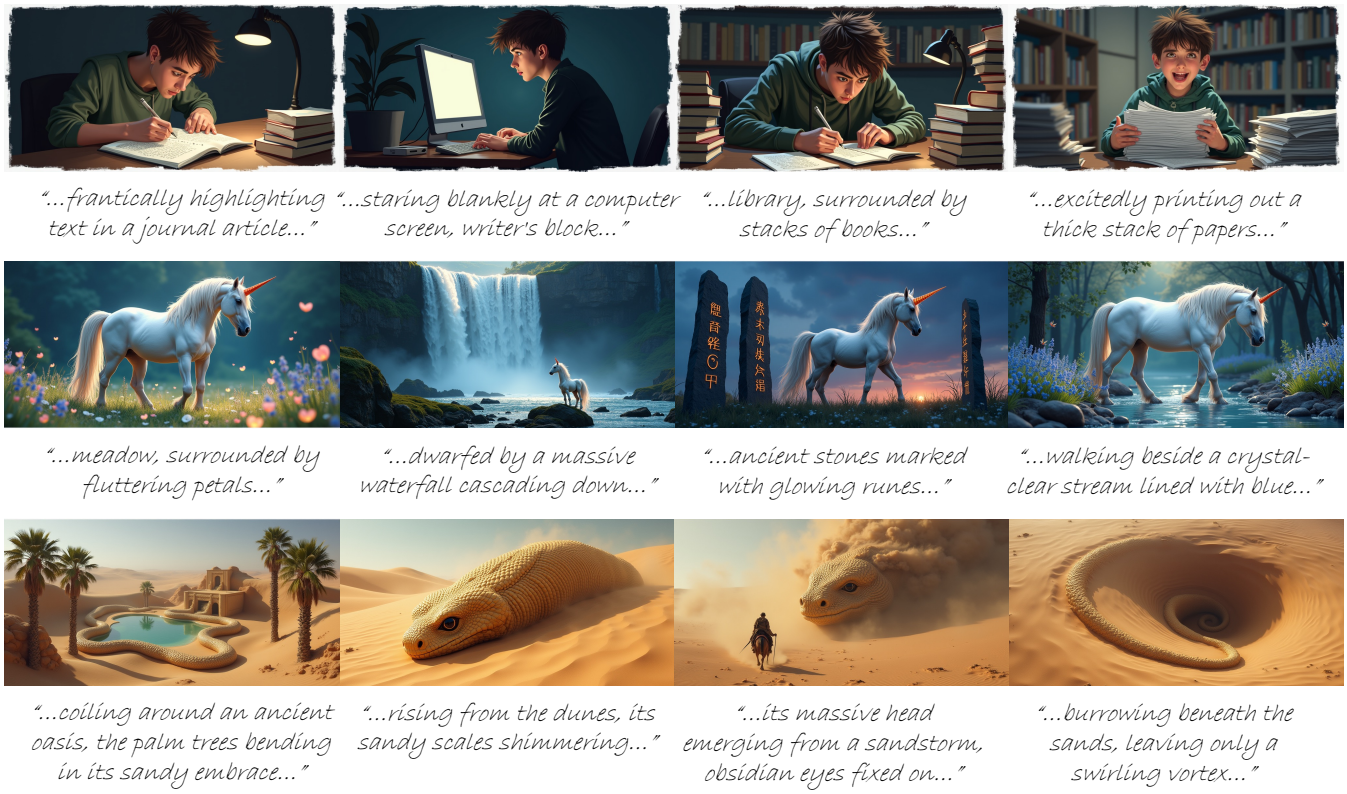

Figure A2: *Additional storyboards generated by our method.*

(2) Across the seven panels, include both together moments and solo moments: exactly five of the seven panels should show both protagonists together, and exactly two should be solo (one panel featuring only character A, and one panel featuring only character B). Vary poses, scale, and camera angle.
(3) Ensure evolving backgrounds and environmental interaction (shifting light/weather, discoveries, traversal).
(4) Provide seven distinct panel descriptions, each a single visually striking moment that highlights teamwork, contrast, or complementary skills.
Return: (a) the full story (one or two short paragraphs) and (b) a list of exactly 7 storyboard panel prompts.

Example Output (Character & Companion): ***Guardians of the Starlit Grove*** **Story.** An elf and her wolf companion crossed into the Starlit Grove beneath constellations bright enough to cast shadows. Among runes carved into a living, glowing tree, the elf traced lines that sang faintly under her fingers; elsewhere, the wolf stood alone, muzzle lifted toward a floating moonstone orb, summoning echoes from the canopy. Together they sprinted through bioluminescent mushrooms, sparks rising in their wake, then discovered a hidden chamber coiled beneath ancient roots where the air smelled of rain and crystal. At a river of stardust they leapt as comet–petals drifted around them. Finally they rested beneath a crystalline vault while the constellations above slowly rearranged, spelling out a promise they would keep.

Storyboard panel prompts.

1. An elf and her wolf companion entering the Starlit Grove under a sky full of constellations
2. An elf alone examining the runes carved into a glowing tree
3. A wolf alone howling at a floating moonstone orb
4. An elf and her wolf racing through a field of bioluminescent mushrooms
5. An elf and her wolf discovering a secret grove chamber hidden beneath ancient roots
6. An elf and her wolf leaping across a river of stardust as comet–petals drift around them
7. An elf and her wolf resting beneath a crystal canopy while constellations slowly rearrange overhead

A.2. Ethics and Diversity Considerations

We designed the benchmark to avoid encoding explicit demographic attributes. Prompts *do not* specify race, ethnicity, or real–world nationalities; they avoid stereotyped roles; and they include a mix of human, animal and fantasy settings, varied body presentations, and roles across genders. Nonetheless, because prompts are generated by a large language model, residual biases in phrasing or depiction may persist. We therefore (i) performed a light manual filtering pass to remove obviously problematic items, (ii)

diversified settings and occupations across entries, and (iii) welcome community feedback and will correct or remove entries that are flagged. The benchmark is intended for research on visual storytelling and character consistency, not for demographic classification or identity inference.

Evaluation on the Rich Storyboard Benchmark. For fair comparison, *all* baselines are evaluated on the *same* set of scene prompts: each benchmark story provides 7 single–panel prompts (one per panel). Our method, however, consumes LPA two–panel prompts. To keep prompt content identical while enabling LPA, we deterministically convert the 7 standard prompts into 6 LPA pairs as follows: (i) designate the first prompt as the *reference*, (ii) pair it as the top sub–panel with each of the remaining six prompts as bottoms (panels 2–7), and (iii) use the "the exact same [character(s)]" phrasing to preserve identity. Thus, the reference panel doubles as panel 1, and each LPA pair corresponds uniquely to one of the original scene prompts (panels 2–7).

This protocol guarantees prompt alignment across methods: competitors run on the original 7 single–panel prompts, while our method runs on the 6 derived LPA pairs whose bottoms are content-matched to panels 2–7. Consequently, results are comparable without prompt mismatch. (Note: although the LLM Director can generate LPA pairs directly from the full story, for benchmark reporting we use the standardized conversion above to ensure consistent evaluation across all approaches.)

B. Implementation Details: Inference Prompts for LPA

We implement our method using the open-source Flux diffusion model [Bla24]. All experiments are conducted on the pre-trained Flux.1-dev model without additional training or finetuning.

B.1. LLM Director: Converting a Story into LPA Two–Panel Prompts

Given a full story text. We use an *LLM Director* to produce (i) a single *reference panel prompt* that clearly depicts the persistent protagonist(s) and their signature attributes, and (ii) a set of two–panel LPA prompts of the form "A storyboard of [REFERENCE] (top) and the exact same [PROTAGONIST(S)] [TARGET] (bottom)". The phrase *the exact same* refers strictly to character identity (appearance/clothing/props) and *not* to style, camera, or background.

Example LLM Director LPA Prompt (inference, 4 panels). Given a full story, select the 4 most visually important moments and describe each as a concise storyboard panel prompt. Do not use names; instead, describe characters directly by appearance, clothing, or role. Ensure descriptors are consistent across all panels so the same protagonist(s) can be recognized.

Next, identify the clearest panel as the *reference panel*. Construct LPA two–panel prompts by pairing the reference panel (top) with each of the other 3 panels (bottom), phrased as: "A storyboard of [REFERENCE] (top) and the exact same [CHARACTER(S)] [TARGET] (bottom)."

Example. For a story about a sailor who journeys across the sea and encounters changing skies and waters, the LPA prompts could be:

- *Reference panel:* a sailor in a wool cap and dark coat standing at the bow of a wooden ship in golden morning fog
- *LPA pairs:*
  1. A storyboard of a sailor in a wool cap and dark coat standing at the bow of a wooden ship in golden morning fog (top) and the exact same sailor sitting on a crate beneath coiled ropes as gulls circle overhead (bottom)
  2. A storyboard of a sailor in a wool cap and dark coat standing at the bow of a wooden ship in golden morning fog (top) and the exact same sailor standing on the deck as the ship approaches a massive stone arch (bottom)
  3. A storyboard of a sailor in a wool cap and dark coat standing at the bow of a wooden ship in golden morning fog (top) and the exact same sailor resting beside the ship's steering wheel, moonlight shimmering on the sea (bottom)

Reference Panel Selection. As described in the main paper, our pipeline requires a single *reference panel* to serve as the source of character identity features. In practice, we reuse one of the storyboard panels that already includes all the main characters–typically the first or second scene in the sequence. We found that this approach produces results as good as or better than using a separately rendered reference panel, while also saving inference time and VRAM.

C. Implementation Details: Hyperparameters & Pseudocode

RAVM Details. We apply Reciprocal Attention Value Mixing (RAVM) at inference time using the following configuration:

- We run inference with classifier-free guidance of 3.5 and 28 denoising steps.
- RAVM is applied to all 38 dual-stream transformer blocks in Flux's denoising network.
- RAVM is applied prior to RoPE
- We use a mixing parameter $\lambda = 0.5$ to control the blend between source and target token values.
- The reciprocal attention maps are smoothed using exponential decay with a momentum of 0.8.

Our full pipeline is training-free, fast to run, and requires no architectural modification to the underlying diffusion transformer. All interventions are performed at the attention value level during sampling.

C.1. Pseudocode for LPA and RAVM

Explanation (LPA). Each scene i is conditioned with a *two–panel prompt* that concatenates the shared reference description on top and the scene–specific description on bottom; these prompts are passed to every denoising block as conditioning. During sampling (from $t{=}T$ down to 1), after each transformer block we *overwrite* the top (reference) half of every sample with the top half of the anchor panel (the first batch element). This keeps identity/appearance cues synchronized across the batch while allowing the bottom halves to evolve according to their own scene prompts. At the end

Algorithm 1 Latent Panel Anchoring (LPA)

Inputs:

- Shared *reference prompt* π_{ref} and per–scene *scene prompts* $\{\pi^{(i)}_{\text{scene}}\}_{i=1}^{n}$.
- Two–panel latent tokens $\{\mathbf{z}^{(i)}\}_{i=1}^{n}$, each split into top (reference) and bottom (scene) halves: $\mathbf{z}^{(i)} = [\mathbf{R}^{(i)}; \mathbf{p}^{(i)}]$.

Output: Bottom panels $\{\hat{\mathbf{p}}^{(i)}\}_{i=1}^{n}$.

1: **Form two–panel prompts:** for each scene i, define $\pi^{(i)} =$ "A storyboard of π_{ref} (top) and $\pi^{(i)}_{\text{scene}}$ (bottom)"

2: **for** $t = T$ **down to** 1 **do** ▷ reverse–time denoising

3: **for** $b = 1$ to B **do** ▷ iterate transformer blocks

4: Denoise batch through block b at step t:

$$\{\mathbf{z}'^{(i)}\} \leftarrow \text{DiTBlock}_b\big(\{\mathbf{z}^{(i)}\}, t, \{\pi^{(i)}\}\big)$$

5: **Anchor top halves: for all** $i = 2, \ldots, n$**, set**

$$\mathbf{R}'^{(i)} \leftarrow \mathbf{R}'^{(1)} \qquad \mathbf{z}'^{(i)} \leftarrow [\mathbf{R}'^{(i)}; \mathbf{p}'^{(i)}]$$
$$\{\mathbf{z}^{(i)}\} \leftarrow \{\mathbf{z}'^{(i)}\}$$

6: Decode each $\mathbf{z}^{(i)}$ and crop the bottom sub–panel $\hat{\mathbf{p}}^{(i)}$.

we decode and retain only the bottom sub–panels for the storyboard.

Algorithm 2 Reciprocal Attention Value Mixing (RAVM)

Notation: batch N, tokens per panel P, feature dim d.

Inputs: concatenated tokens $\mathbf{x} = [\mathbf{x}_{\text{top}}; \mathbf{x}_{\text{bot}}] \in \mathbb{R}^{N\times(2P)\times d}$; projections $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$; mixing weight $\lambda \in [0,1]$; reciprocal–attention history $\bar{\mathbf{M}} \in \mathbb{R}^{N\times P\times P}$ with decay $\beta \in (0,1)$.

Output: Updated values $\mathbf{V}$, then SDPA with $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$.

1: $\mathbf{A} \leftarrow \text{softmax}\big(\mathbf{Q}\mathbf{K}^{\top}/\sqrt{d_k}\big)$ $\quad \in \mathbb{R}^{N\times(2P)\times(2P)}$

2: Extract cross–panel attention maps:

$$\mathbf{A}_{top\to bot} = \mathbf{A}[:, 1{:}P, P{:}2P], \quad \mathbf{A}_{bot\to top} = \mathbf{A}[:, P{:}2P, 1{:}P]$$

3: **Reciprocal attention scores (elementwise minimum):**

$$\mathbf{M} \leftarrow \min\big(\mathbf{A}_{bot\to top}, \mathbf{A}^{\top}_{top\to bot}\big) \quad \in \mathbb{R}^{N\times P\times P}$$

4: **Update history:** $\bar{\mathbf{M}} \leftarrow \beta\bar{\mathbf{M}} + (1-\beta)\,\mathbf{M}$

5: **Sparse bottom–token mask:** threshold $\bar{\mathbf{M}}$ with Otsu *per batch* (on columns) and clean small components to obtain $\mathbf{B} \in \{0,1\}^{N\times P}$ ▷ $\mathbf{B}[n,v]$=1 iff bottom token v in sample n has sufficient reciprocal support

6: **Match top partners (vectorized, per sample):**

$$\text{top\_idx} \leftarrow \text{argmax}_u\ \bar{\mathbf{M}}[:, u, :] \quad \in \mathbb{N}^{N\times P}$$

7: **Gather matched top values:**

$$\mathbf{V}_{\text{mix}}[n, v, :] \leftarrow \mathbf{V}_{\text{top}}\big[n, \text{top\_idx}[n, v], :\big] \quad \in \mathbb{R}^{N\times P\times d}$$

8: **Masked linear blend:**

$$\mathbf{V}_{\text{bot}} \leftarrow (1-\mathbf{B}) \odot \mathbf{V}_{\text{bot}} + \mathbf{B} \odot \big(\lambda\,\mathbf{V}_{\text{bot}} + (1-\lambda)\,\mathbf{V}_{\text{mix}}\big)$$

9: Run SDPA with unchanged $(\mathbf{Q}, \mathbf{K})$ and updated $\mathbf{V}$.

Explanation (RAVM). We compute attention over the concatenated top/bottom tokens and form a *reciprocal* map by the elementwise minimum of the two cross–panel blocks, keeping only mutually strong links. We maintain a *reciprocal attention history* $\bar{\mathbf{M}}$ via exponential smoothing across diffusion *steps and blocks*, with decay β; **at the very first denoising step (before any transformer block) we initialize $\bar{\mathbf{M}}$ to zero**. To decide which bottom tokens to update, we apply Otsu thresholding to $\bar{\mathbf{M}}$ columnwise (per sample) and remove tiny components, yielding a binary mask $\mathbf{B} \in \{0,1\}^{N\times P}$. For each sample n and bottom token v, we pick its best top partner by a columnwise arg max on $\bar{\mathbf{M}}[n,:,v]$, gather the corresponding top value vector into $\mathbf{V}_{\text{mix}}[n,v,:]$, and perform a single masked linear blend $(1-\mathbf{B}) \odot \mathbf{V}_{\text{bot}} + \mathbf{B} \odot (\lambda\mathbf{V}_{\text{bot}} + (1-\lambda)\mathbf{V}_{\text{mix}})$. Here $\mathbf{B}$ is broadcast along the feature dimension d. Because queries and keys are unchanged, spatial layout and attention routing remain intact while appearance features propagate via the values.

D. DS-500 Evaluation

To assess generalization beyond our benchmark, we also evaluate on DS-500 [HYT*24], a storyboard dataset with shorter prompts and minimal scene evolution. While not designed to test layout or narrative expressivity, DS-500 serves as a useful baseline for identity coherence and basic prompt alignment.

Method	CLIP-T (↑)	DreamSim (↑)
DreamStory [HYT*24]	**0.3779**	0.6714
Story2Board (ours)	0.3723	**0.7018**

Table A1: *DS-500 evaluation results. DreamStory's scores are reported directly from their paper [HYT*24]. Our method achieves competitive CLIP-T alignment while outperforming DreamStory in identity consistency (DreamSim).*

As shown in Table A1, our method achieves comparable prompt alignment (CLIP-T) [RKH*21] and higher identity consistency (DreamSim [FTS*23]) relative to DreamStory [HYT*24]. This supports the broader applicability of our approach across datasets with varying narrative complexity.

E. Ablation: Reciprocal Mixing Neighborhood Size

Setup. Our default RAVM selects, for each bottom token v, the *single* top token $u^{\star} = \arg\max_u \bar{M}[u,v]$ and mixes only that value vector. We ablate this choice by allowing a *top-k* neighborhood: for each v, we take the k highest–scoring top tokens under $\bar{M}$ and mix their average values (V) into v. All other settings, schedules, and evaluation protocol remain unchanged.

Metrics. We report character consistency (DreamSim), prompt alignment (VQAScore), and our Scene Diversity metric. Figures below show the pairs omitted from the main text.

Findings. (1) k=1 **(argmax) yields the best identity fidelity.** Increasing k consistently *reduces* DreamSim, indicating that averaging multiple sources softens distinctive appearance cues and introduces cross–region leakage. (2) **Prompt alignment shows no compensating gains.** VQAScore remains essentially unchanged across k, so the loss in identity is not offset by better grounding (Fig. A3). (3) **Scene diversity does not improve with larger k.** We observe negligible or no increase in our Scene Diversity metric as k grows

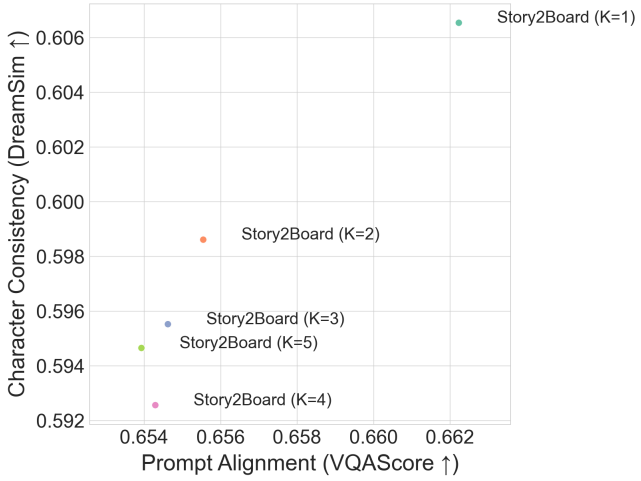

Figure A3: ***Prompt Alignment (VQAScore ↑) vs. Character Consistency (DreamSim ↑)*** *for top-k mixing. Each point is Story2Board with* $k \in \{1,2,3,4,5\}$.

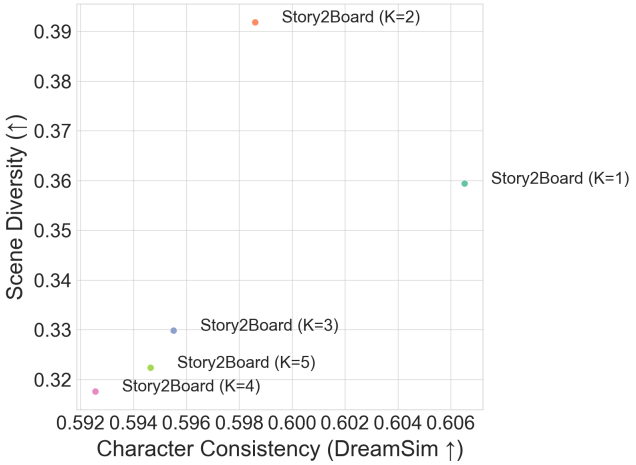

Figure A4: ***Character Consistency (DreamSim ↑) vs. Scene Diversity*** *(↑) for top-k mixing. Each point is Story2Board with* $k \in \{1,2,3,4,5\}$.

(Fig. A4). This suggests that compositional variety is already governed by queries/keys and the denoising dynamics; broadening the value–mixing neighborhood mainly blurs identity rather than creating new layouts.

Discussion. Top-k mixing intuitively aggregates appearance from multiple candidate correspondences, but in practice those correspondences often map to adjacent or partially overlapping regions, leading to value–space averaging that erodes fine–grained details (hair, textures, accessories). Since RAVM purposefully leaves (Q,K) unchanged (layout is preserved there), expanding the value neighborhood cannot inject new spatial diversity and therefore yields little benefit.

Conclusion. Given weaker identity fidelity, no measurable prompt–grounding gains, no improvement in diversity, and extra compute from gathering k sources, we adopt k=1 in all main results.

F. In Depth: Reciprocal Attention Value Mixing

F.1. Extended explanation of Figure 4

What the diagram conveys. Figure 4 visualizes how *Reciprocal Attention Value Mixing (RAVM)* identifies semantically corresponding tokens across the two stacked sub-panels (reference on top, target on bottom) at an intermediate diffusion step. The red and green markers indicate a specific semantic part (the hand) in the top and bottom sub-panels, respectively. The two heatmaps on the right show, for that same denoising step, the *reciprocal attention* scores between (i) *every token in the top panel* and the *selected bottom token* (top-right map), and (ii) *every token in the bottom panel* and the *selected top token* (bottom-right map). In both views, the strongest response localizes at the corresponding hand token, illustrating that RAVM surfaces aligned token pairs suitable for value mixing, thereby reinforcing appearance consistency without altering spatial composition.

F.2. How we generated the diagram (step-by-step).

1. *Run with LPA and RAVM enabled.* We perform standard inference with our pipeline: prompts are two-panel (top reference, bottom scene), Latent Panel Anchoring anchors the top halves across the batch, and RAVM is active on a range of transformer blocks and diffusion steps (as in the main experiments). The figure uses *denoising step* t=12 out of T=28 steps.
2. *Hook into a DiT block.* At step t=12, we instrument one of the RAVM-active DiT blocks (a mid-level block) and record the attention tensors *before* the value mixing assignment is applied (to avoid circularity in the visualization).
3. *Extract multi-head attention.* Let B be the batch size (here B=1 for visualization), H the number of heads, D_h the per-head dimension, and T the token length for this block. After the standard linear projections,

$$Q \in \mathbb{R}^{B\times H\times T\times D_h}, \quad K \in \mathbb{R}^{B\times H\times T\times D_h}, \quad V \in \mathbb{R}^{B\times H\times T\times D_h},$$

attention per head is

$$A^{(h)} = \mathrm{softmax}\left(\frac{Q^{(h)}(K^{(h)})^\top}{\sqrt{D_h}}\right) \in \mathbb{R}^{T\times T} \quad (h = 1,\ldots,H).$$

We average over heads to obtain a single score matrix $\bar{A} \in \mathbb{R}^{T\times T}$.

4. *Index the prompt/image regions.* Tokens are ordered as

$$\underbrace{\text{prompt}}_{N_p} \mid \underbrace{\text{top image}}_{P} \mid \underbrace{\text{bottom image}}_{P},$$

so $T = N_p + 2P$. In our setup each sub-panel has $P = H_{\text{grid}} \times W_{\text{grid}}$ tokens; we use H_{grid}=32, W_{grid}=64 (thus P=2048), matching the latent token grid used by the DiT. We slice the cross-panel blocks

$$A_{tb} = \bar{A}[N_p{:}N_p{+}P,\ N_p{+}P{:}N_p{+}2P] \in \mathbb{R}^{P\times P},$$

$$A_{bt} = \bar{A}[N_p{+}P{:}N_p{+}2P,\ N_p{:}N_p{+}P] \in \mathbb{R}^{P\times P},$$

corresponding to *top→bottom* and *bottom→top* attention.

5. *Compute reciprocal attention.* The reciprocal attention matrix is

$$M = \min\left(A_{bt},\, A_{tb}^{\top}\right) \in \mathbb{R}^{P\times P} \quad \text{(elementwise min)}.$$

For Figure 4 we display the EMA-smoothed map $\bar{M}$ at block/step $(b,t) = (\text{mid}, 12)$, i.e., the same quantity RAVM uses for mixing.

6. *Select the two anchor tokens.* We pick one token index $u_{\text{red}} \in \{1,\dots,P\}$ on the top sub-panel (centered on the red circle) and one token $v_{\text{green}} \in \{1,\dots,P\}$ on the bottom sub-panel (centered on the green circle). Selection is done by mapping the 2D circle location to the nearest token in the 32×64 grid.

7. *Form the two heatmaps.* (i) *Top-right map:* column v_{green} of M gives $\{M[u, v_{\text{green}}]\}_{u=1}^{P}$, the reciprocal attention of every *top* token to the selected *bottom* token. We reshape this P-vector to $(32, 64)$ and display it over the top panel. (ii) *Bottom-right map:* row u_{red} of M gives $\{M[u_{\text{red}}, v]\}_{v=1}^{P}$, the reciprocal attention of every *bottom* token to the selected *top* token. We reshape to $(32, 64)$ and display it over the bottom panel.

Tensor shapes at a glance.

Q, K, V	$\in \mathbb{R}^{B\times H\times T\times D_h}$
$\bar{A}$	$\in \mathbb{R}^{T\times T}$ (head-averaged)
A_{tb}, A_{bt}, M	$\in \mathbb{R}^{P\times P}$ $(P = 32\times 64 = 2048)$
slice $M[:, v_{\text{green}}]$, $M[u_{\text{red}}, :]$	$\in \mathbb{R}^{P} \xrightarrow{\text{reshape}} \mathbb{R}^{32\times 64}$

Interpretation. The brightest responses in both heatmaps align with the hand token in the opposite panel, indicating a strong bidirectional (reciprocal) linkage between the two tokens. This is precisely the signal RAVM exploits: for selected bottom tokens v, it blends their value vectors with those of their maximally aligned top tokens $u^{\star} = \arg\max_u M[u, v]$, thereby refining appearance where semantic correspondence is strongest, while leaving attention routing and layout (keys/queries) unchanged at the point of application.

G. Cross-panel Attention Diagnostic

We measure how much attention mass bottom-panel tokens assign to the reference (top) panel, separately for foreground vs. background regions. For each bottom token i, we compute $\alpha_i = \sum_{j\in \text{Top}} A_{ij}$, where A denotes the post-softmax attention weights and Top are the indices of top-panel tokens. We then average α_i over bottom tokens inside the foreground mask (used by RAVM) and over tokens outside it (background), and report mean±std over prompts. We measure how much attention mass bottom-panel tokens assign to the reference (top) panel, separately for foreground vs. background regions. Here, *foreground* denotes the subset of bottom-panel tokens selected by our character mask (the same tokens that participate in RAVM value mixing), and *background* denotes the remaining bottom-panel tokens. For each bottom token i, we compute $\alpha_i = \sum_{j\in \text{Top}} A_{ij}$, where A denotes the post-softmax attention weights and Top are the indices of top-panel tokens. We then average α_i over foreground tokens and over background tokens, and report mean±std over prompts. We expect foreground tokens to exhibit higher cross-panel attention than background tokens, since identity-relevant regions (e.g., face/clothing) form stronger links to the reference. Observing this gap supports our intuition that RAVM primarily couples character appearance across panels while leaving background composition comparatively unconstrained.

Table A2: *Cross-panel attention diagnostic (foreground vs. background). We report the average fraction of attention mass from* bottom-panel *tokens to the reference (*top*) panel, for background tokens vs. foreground (character/mixed) tokens (mean±std over 700 prompts).*

	Background	Foreground
Bottom→Top attention	0.356 ± 0.024	0.399 ± 0.027

H. User Study Details

To evaluate the perceived quality of generated storyboards, we conducted a large-scale user study on Amazon Mechanical Turk (MTurk) [Ama25] using all 100 stories from our Rich Storyboard Benchmark. We used the first four storyboard panels from each story as rendered by our method and one competing baseline. The resulting pairs were shown to participants side-by-side.

Study Design. We ran five separate studies, each targeting a specific evaluation criterion:

1. **Overall Preference**
2. **Prompt Alignment**
3. **Character Consistency**
4. **Background Richness**
5. **Scene Diversity**

To ensure fair coverage under a fixed budget, we evaluate each baseline on a random subset of 20 stories drawn from the 100-story benchmark. For every sampled story, we run a head-to-head A/B comparison between that baseline and *Story2Board*.

Participant Selection. To ensure high-quality responses, we restricted participation to workers located in English-speaking countries–specifically the United States, United Kingdom, Canada, and Australia. We further filtered for workers with a lifetime task approval rate above 98%, prioritizing reliable and experienced annotators.

Interface and Task. For each evaluation criterion, participants were shown a sequence of trials. Each trial displayed two storyboards (A and B) generated from the same story prompt–one from Story2Board and one from a competing model. Participants were asked to select the storyboard that best satisfied the target criterion. Model names and ordering were not shown.

Each trial presented all four storyboard panels per model, along with their associated captions. Image layouts were standardized and left–right positioning was randomized. Each trial was rated by 3 unique workers.

Instructions to Participants. Participants were given the following instructions at the start of each task. The only variation across studies was the criterion description, shown in bold.

For other criteria, the bolded instruction was replaced accordingly. For instance, for *Prompt Alignment*, participants were asked to choose the version that more accurately matched the text descriptions; for *Scene Diversity*, they were asked to consider how much variety was present across the panels in terms of framing, layout, and setting. Screenshots are presented in Figure A5.

For criteria that required more subjective interpretation–such as *Scene Diversity* and *Background Richness*–participants were also shown example pairs of good and poor storyboards illustrating the concept, drawn from baseline methods and distinct stories not used in the evaluation.

Result Aggregation. Participant responses were aggregated across all trials per criterion to compute win rates. These results are summarized in Figure 8 in the main paper. Our method received the highest overall preference scores, winning the majority of pairwise comparisons in the "Overall Preference" category. This indicates that when participants considered the storyboards as a whole, they consistently favored our approach over all competitors.

At the same time, the results highlight specific trade-offs across individual evaluation criteria. OminiControl achieved stronger scores in prompt alignment, background richness, and scene diversity, likely benefiting from its encoder-based layout conditioning. Meanwhile, IC-LoRA (Storyboards) and StoryDiffusion were slightly favored for character consistency, reflecting their targeted emphasis on identity preservation. In contrast, our method's use of soft, token-level guidance enables greater flexibility in layout and framing–traits that may account for its overall appeal despite falling behind in some focused categories.

I. Additional Metric Visualizations

For completeness, we include all pairwise plots among the three evaluation axes used in the paper–*prompt alignment*, *character consistency*, and *scene diversity*–that were omitted from the main text (Figures A6-A15). Prompt alignment is reported as **VQAScore** [LPL*24] and **CLIP Max**. For CLIP Max we adapt to rectangular outputs by computing CLIP [RKH*21] image–text similarity on (i) a centered square crop and (ii) a square *letterboxed* (padded) version of the image, and taking the maximum of the two. Character consistency is measured by **DreamSim** [FTS*23], **DINOv2** [LZR*24], **CLIP**, and **LPIPS** [ZIE*18] (*lower is better*). **Scene Diversity** is our proposed metric.

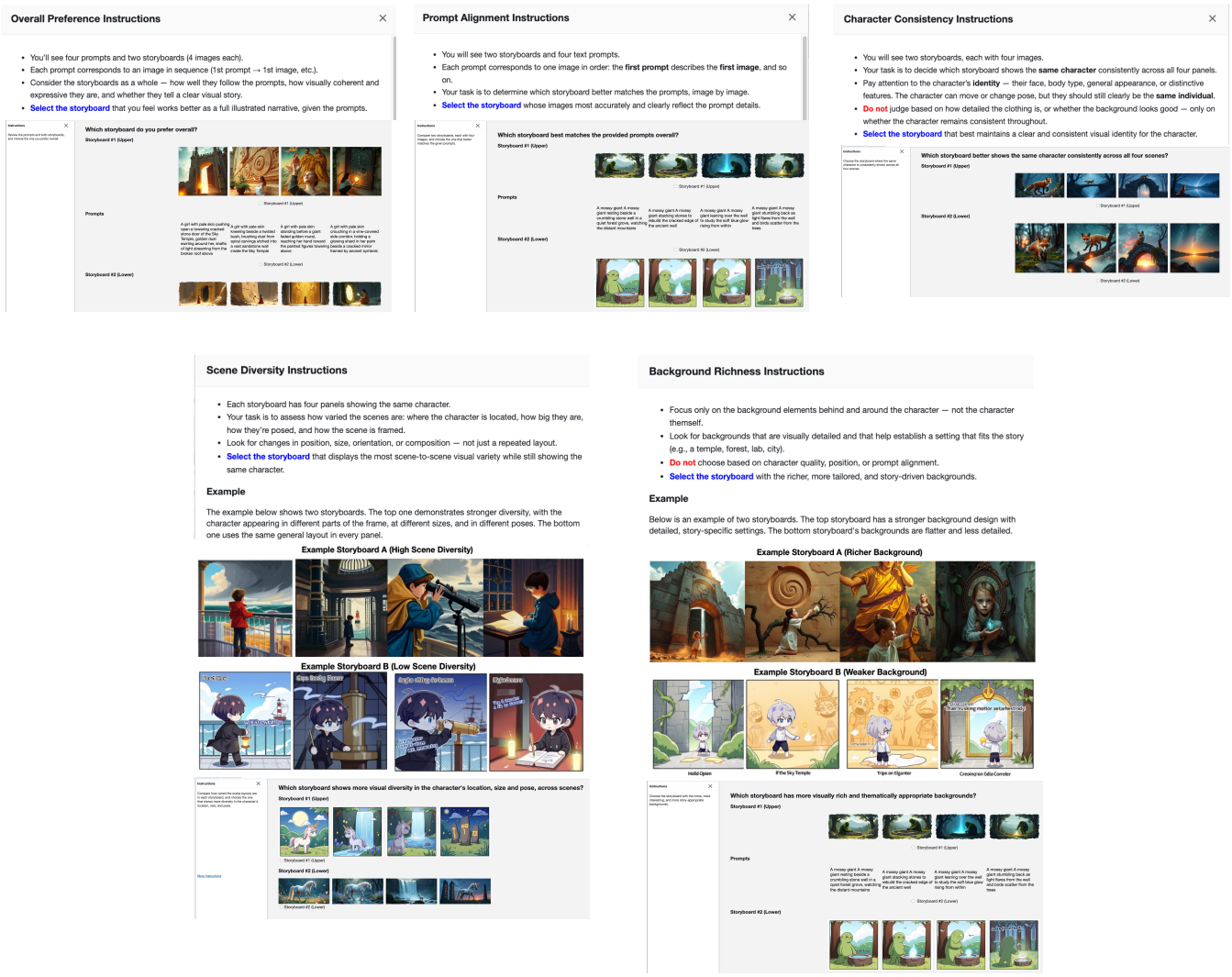

Figure A5: *User Study Instructions. We provide the complete instructions for the user study we conducted using Amazon Mechanical Turk (AMT) to compare our method with each baseline.*

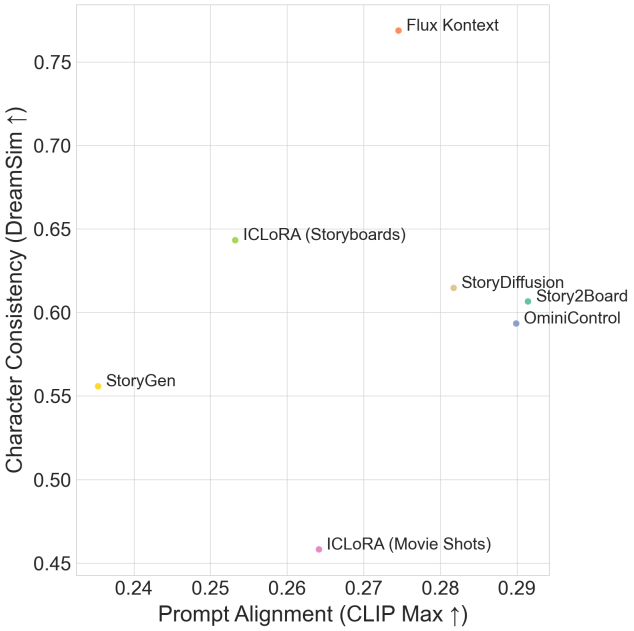

Figure A6: ***Prompt Alignment (CLIP Max ↑) vs. Character Consistency (DreamSim ↑).***

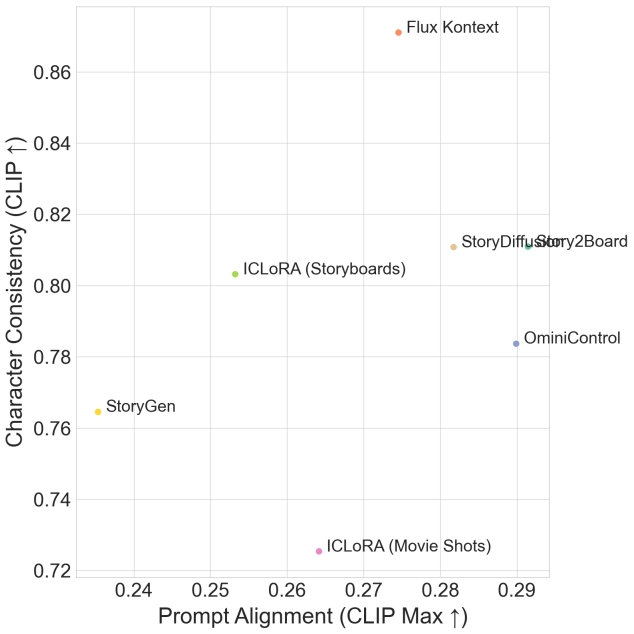

Figure A7: ***Prompt Alignment (CLIP Max ↑) vs. Character Consistency (CLIP ↑).***

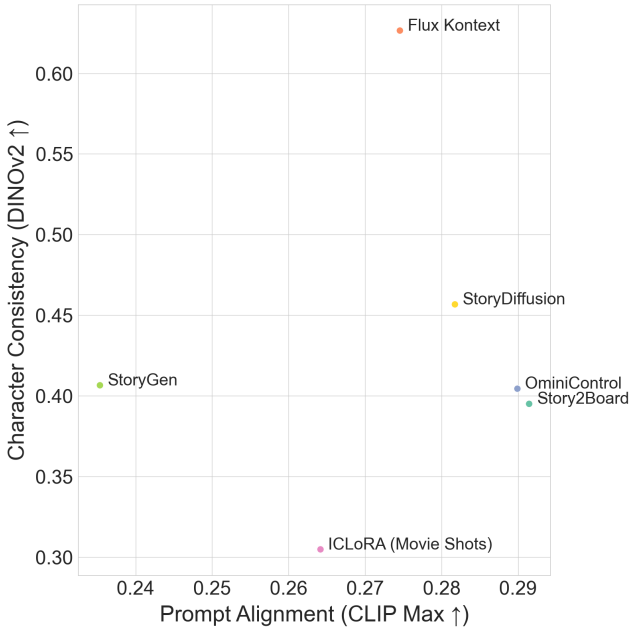

Figure A8: ***Prompt Alignment (CLIP Max ↑) vs. Character Consistency (DINOv2 ↑).***

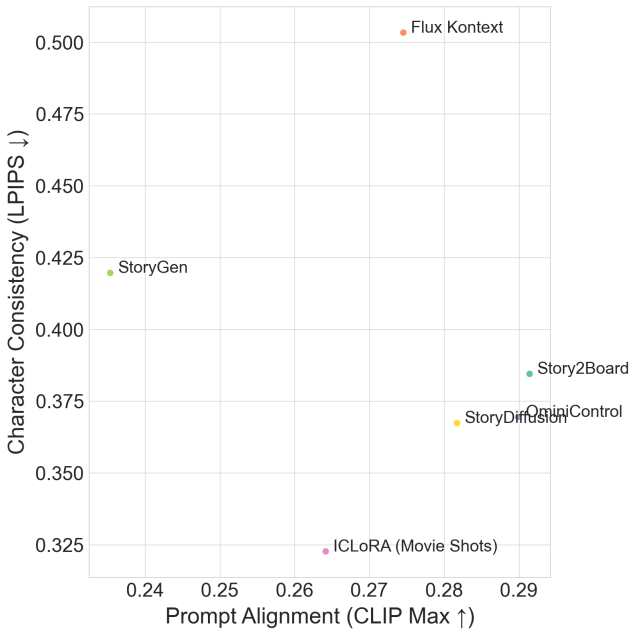

Figure A9: ***Prompt Alignment (CLIP Max ↑) vs. Character Consistency (LPIPS ↓).*** *LPIPS is lower-better.*

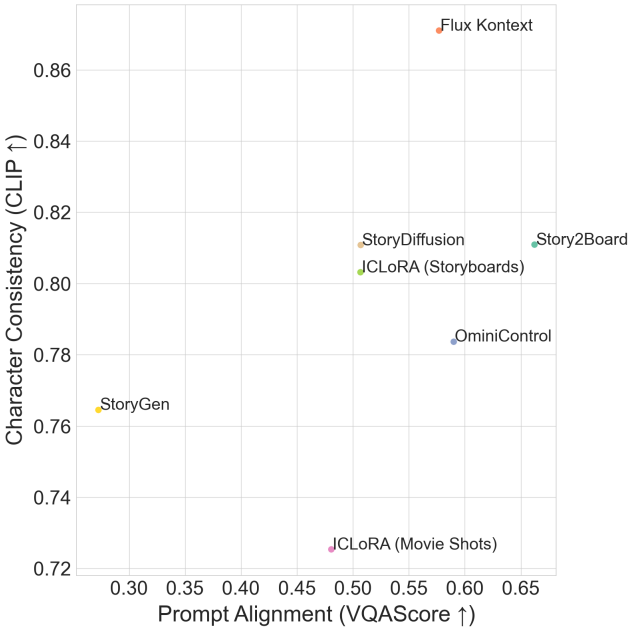

Figure A10: ***Prompt Alignment (VQAScore ↑) vs. Character Consistency (CLIP ↑).***

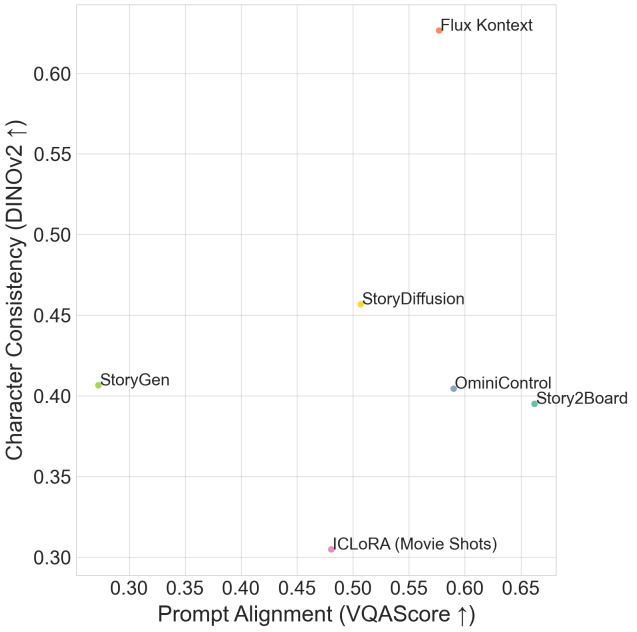

Figure A11: ***Prompt Alignment (VQAScore ↑) vs. Character Consistency (DINOv2 ↑).***

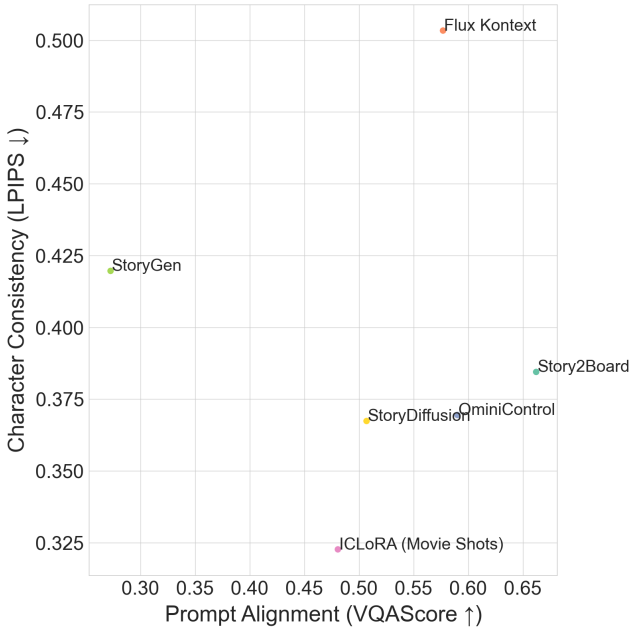

Figure A12: ***Prompt Alignment (VQAScore ↑) vs. Character Consistency (LPIPS ↓).*** *LPIPS is lower-better.*

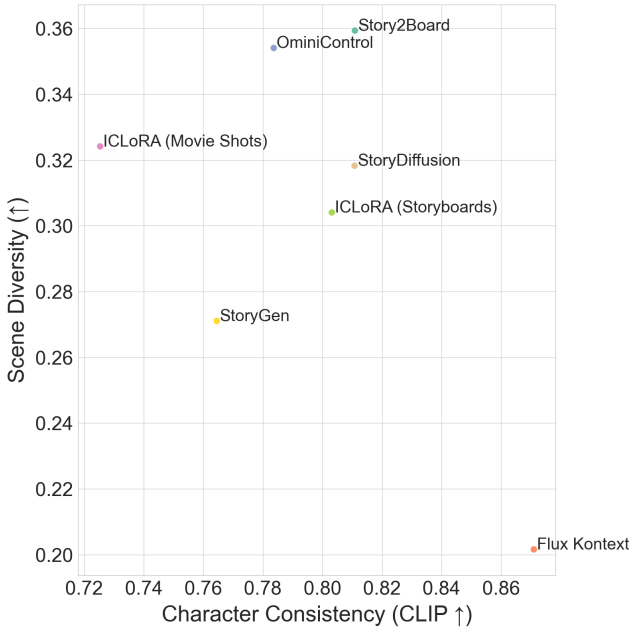

Figure A13: ***Character Consistency (CLIP ↑) vs. Scene Diversity (↑).***

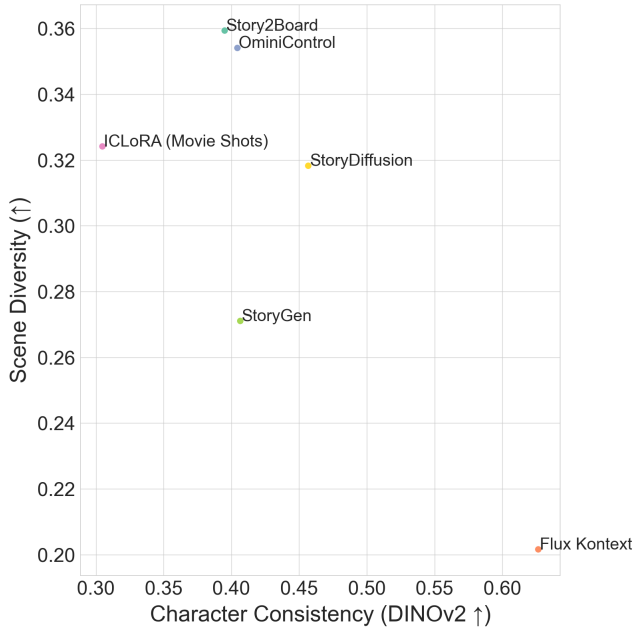

Figure A14: ***Character Consistency (DINOv2 ↑) vs. Scene Diversity (↑).***

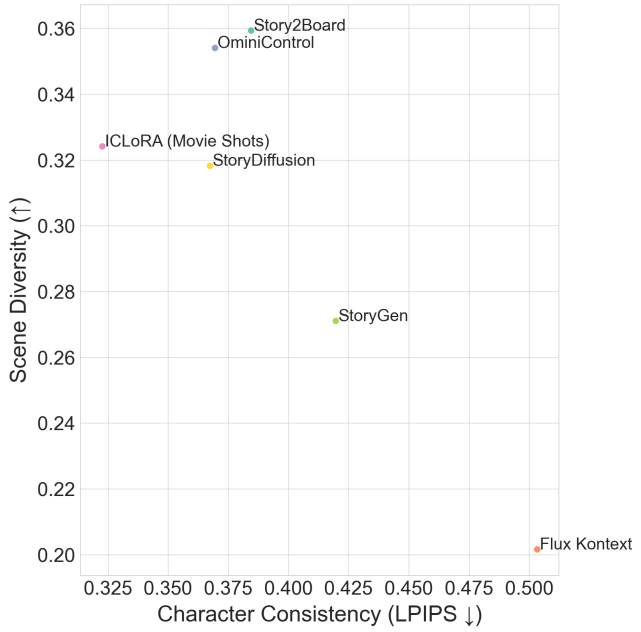

Figure A15: ***Character Consistency (LPIPS ↓) vs. Scene Diversity (↑).*** *LPIPS is lower-better.*

J. Full Story Texts

See Table A3 for examples of complete stories.

Table A3: *Full Story Texts*

Storyboard ID	Full Story Text
Rami the Desert Nomad	Rami, the desert nomad, held his glowing lantern high beside his patient camel, the twilight painting the dunes in shades of blood and gold. The endless sands stretched away into eternity, broken only by whispers of ancient paths. As he pressed forward, Rami passed beneath a ruined sandstone arch, its surface etched by time. Crimson petals blew past in sudden gusts, swirling around him like lost memories, as if the gate itself were exhaling the past. That night, silhouetted against a crescent moon atop a dune ridge, Rami paused. The wind tugged at his cloak while the cold stars wheeled above, casting long shadows across the rippling sand. Seeking shelter, he settled beneath a jagged stone outcrop, kneeling on worn rock. By the glow of his lantern, he unrolled a crumpled map, squinting at the faded lines and markings, unsure of what was memory and what was myth. At dawn, a caravan of dune-moth herders emerged from the haze of a violet dust storm. Rami approached cautiously, negotiating passage through their shifting territory. Strange banners fluttered from their saddles; their moths blinked with luminous eyes. By dusk, he reached a tower of bones–an ancient, impossible spire that pierced the desert sky. Rami climbed its spiraling ramp, each step echoing with forgotten oaths, until he stood at its apex. There, atop a glassy dune, Rami raised his lantern one final time. Its golden glow danced against the wind as twin moons rose behind him. In the distance, tiny signals flickered in reply–other wanderers answering his call across the sands.
Blackpaw in the Celestial Forest	Blackpaw, a shimmering fox of the ancient celestial forest, stepped lightly onto a mossy stone path, the twilight trees arching high above him like a cathedral ceiling. With a flick of his glowing tail, he bounded across a fallen tree stretched precariously over a mist-shrouded ravine that gleamed faintly with constellations reflected in the fog below. Perched atop a broken archway of ancient stone, vines and silver moss hanging around him, Blackpaw gazed out over the glowing forest as twilight deepened. From the edge of a luminous lake mirroring the heavens perfectly, he watched a meteor shower ignite the sky, each fiery streak mirrored twice over. Curling beside a pulsing crystal monolith, he dreamed in the ancient heartbeat of the forest. By morning, the grove was silent but for whispering silver leaves shedding light into the wind, and the fading trace of the fox's gleaming trail.
The Last Astronomer	They called her Dr. Elira Voss, though no one had used her title in years. She was the last custodian of the Skyreach Observatory, a rusting dome perched on the cliffs where stars once spoke to science. Beneath its cracked shutters and wind-scoured walls, Elira still watched the sky–not for data, but for memory. A single tear traced her cheek as a meteor shower flared across the heavens, scattering silver sparks over the dark sea. She stood silently beside a weathered telescope, its brass fittings dulled by time, and turned back to her hand-drawn sky chart. The map was crowded with inked constellations, margins lined with notes and dates only she understood. She moved carefully, peering through the eyepiece and adjusting the scope until a distant galaxy came into view. Her fingers trembled, not from age, but from the echo of another life. On the desk nearby sat a faded photograph of a man in an astronaut's suit–his smile still intact, his absence louder than ever. Later that night, she spotted it: a new star, impossibly bright. Her breath caught. She smiled–not wide, not triumphant, but soft, as if welcoming an old friend. She cranked the observatory's rusted gears, pulling open the cracked dome just in time to follow a teal comet slicing across the sky, its fire washing over ancient machinery like a blessing. And when her hands could do no more, she stepped onto the rooftop and lit a paper lantern. As it rose, its glow joined the glinting trail of satellite beacons. A message. A memory. A promise that she was still watching, still waiting. Still listening.