AttriStory: Fine-grained Attribute Realization for Visual Storytelling with Diffusion Models

Manogna Sreenivas Rohit Kumar Soma Biswas
Indian Institute of Science, Bengaluru, India
{manognas, rohitk, somabiswas}@iisc.ac.in

Figure 1. **Visualization of a story generated from the AttriStory benchmark.** This story of *Ben*, illustrates the dual challenge in visual storytelling: maintaining character consistency across scenes, while realizing fine-grained attributes such as clothing and accessories.

Abstract

Visual storytelling with diffusion models has made impressive strides in maintaining character consistency across narrative scenes. However, a critical gap remains: while these methods ensure a character remains consistent across scenes, they provide no systematic method to ensure if fine-grained attributes such as color and textures of clothing, accessories are faithfully rendered in the generated images. Towards this goal, we introduce AttriStory, a benchmark enabling attribute realization in visual storytelling. We curate 200 multi-scene stories across 10 distinct artistic styles using Large Language Model. Each scene is constructed with detailed attribute specifications to enable rich visual narratives. Further, to address attribute realization, we propose a plug-and-play latent optimization module that operates during early denoising steps, when the model establishes structural and semantic content. We achieve this through AttriLoss objective designed to maximize alignment between the cross-attention maps for desired attribute-object pairs while suppressing spurious associations, guiding models to localize attributes correctly. This approach operates orthogonally to existing consistency mechanisms, integrating seamlessly with current story generation pipelines without requiring architectural modifications. Our experiments demonstrate consistent improvements on incorporating AttriLoss across all baselines. This work positions attribute realization as a distinct, complementary dimension of visual storytelling, alongside character consistency, advancing the field toward fine-grained attribute-controlled story generation.

1. Introduction

Visual storytelling with diffusion models [7, 17, 19, 20, 22, 38] has become a powerful tool for creative industries, enabling artists, animators, and designers to generate coherent story narratives with unprecedented character consistency. Recent advances [18, 32, 40] exemplify significant progress in this direction, leveraging training-free mechanisms to preserve character identity and appearance across scenes through extended self-attention and feature sharing.

Despite these successes, we identify a critical gap in the visual storytelling pipeline: while existing methods [18, 32, 33, 40] prioritize character consistency, fine-grained visual attributes remain underexplored and under-specified in both generation and evaluation. Current approaches focus pri-

marily on basic character identity while largely neglecting details like clothing color, fabric texture, and accessories, which are elements essential for high-quality visual storytelling. This limitation becomes apparent in real-world creative workflows. In professional animation studios or editorial illustration, storytellers demand precise control over visual details. An artist might specify not merely *"a boy playing with a dog"* but rather *"A cartoon illustration of a boy wearing red shirt and blue sneakers playing with a brown corgi"*, where each attribute carries narrative significance. These fine-grained specifications are essential for conveying genre, details of clothing, accessories, particularly in domains like comics and animation where attribute-level fidelity directly impacts storytelling quality.

We define this challenge as the attribute realization problem: ensuring that explicitly specified, fine-grained visual attributes (e.g., *"red shirt", "blue sneakers", "brown corgi"*) are actually realized in generated images. This problem is fundamentally orthogonal to cross-scene character consistency. While identity preservation ensures a character's recognizable features persist across scenes, attribute realization addresses whether detailed cues in the narrative text are correctly bound to visual elements in each frame. Although a model may successfully maintain character consistency, it might fail to render specified fine-grained attributes, despite explicit textual instruction. We address this gap, we make the following key contributions:

- We introduce a benchmark named *AttriStory* designed specifically to evaluate attribute realization in visual storytelling. This extends narrative scenarios with detailed attribute specifications, enabling systematic measurement of whether generated images conform to fine-grained visual descriptions. We curate 200 multi-scene stories across 10 distinct artistic styles using an LLM to generate coherent narratives with accompanying character descriptions, scene-specific actions, and positive and negative attribute-object associations (e.g., positive: *["red", "shirt"]*; negative: *["blue", "shirt"]*). This approach enables scalable benchmark creation by simulating realistic artist workflows where designers specify both narrative content and desired visual attributes.
- We propose *AttriLoss*, a latent optimization objective designed to enhance attribute realization, enabling plu-and-play integration with current storytelling pipelines. This update is done during the early denoising steps when diffusion models establish structural and semantic content including colors and textures. We optimize cross-attention-based Intersection-over-Union (IoU) losses to maximize alignment between attention maps for desired attribute-object pairs while suppressing IoU between undesired attribute-object associations. Importantly, our approach complements rather than replacing existing consistency mechanisms, enabling seamless integration with recent story generation methods [32, 40], requiring no modifications to their core architectures.
- Our experiments on integrating *AttriLoss* with baseline storytelling methods and evaluation on *AttriStory* benchmark demonstrate consistent and significant improvements across all evaluated baselines and artistic styles. This work positions attribute realization as a distinct, complementary dimension of visual storytelling, alongside character consistency, providing a systematic benchmark and a practical method to advance the field toward fine-grained visual narrative generation.

2. Related Work

Visual storytelling with diffusion models sits at the intersection of consistent character generation and text-to-image synthesis. Here, we review recent advances in these areas.

Text-to-Image Generation using Diffusion Models. Recent advances in diffusion models have revolutionized image generation [11, 13, 16, 23, 26, 30, 36, 37, 41]. Denoising Diffusion Probabilistic Models (DDPM) [10] established the foundation for generating high-quality images through iterative denoising. To accelerate generation, Denoising Diffusion Implicit Models (DDIM) [31] introduced a deterministic sampling procedure that significantly reduces inference steps while maintaining quality. Latent Diffusion Models [27], exemplified by Stable Diffusion [24], extended this paradigm to operate in compressed latent space, improving computational efficiency.

A key insight that enables controllable generation is the role of cross-attention mechanisms in diffusion models. Cross-attention layers bind textual tokens to spatial regions in the image through the CLIP text encoder [25], making them well-suited for editing. Prompt-to-Prompt [8] for image editing, leveraged this observation to enable training-free editing by modifying cross-attention maps during the denoising process. This work also demonstrated that early denoising steps establish coarse structure and semantic content (including colors and textures), while later steps refine fine details. Subsequent works [2, 12] have adopted similar mechanisms for personalization objectives, establishing cross-attention manipulation as a principled approach for guiding diffusion model generation.

Visual Storytelling and Character Consistency. Character consistency across multiple scenes represents a critical challenge in visual storytelling applications. Recent research has developed two main paradigms: personalization-based approaches [1, 2, 6, 28, 29, 34, 35] and training-free self-attention based consistency mechanisms [4, 18, 32, 40].

Early personalization methods fine-tune diffusion models for specific identities. DreamBooth [28] fine-tunes the full model, while Textual Inversion [6] learns unique embeddings. Custom Diffusion [12] reduces the compute overhead by fine-tuning only cross-attention projections, high-

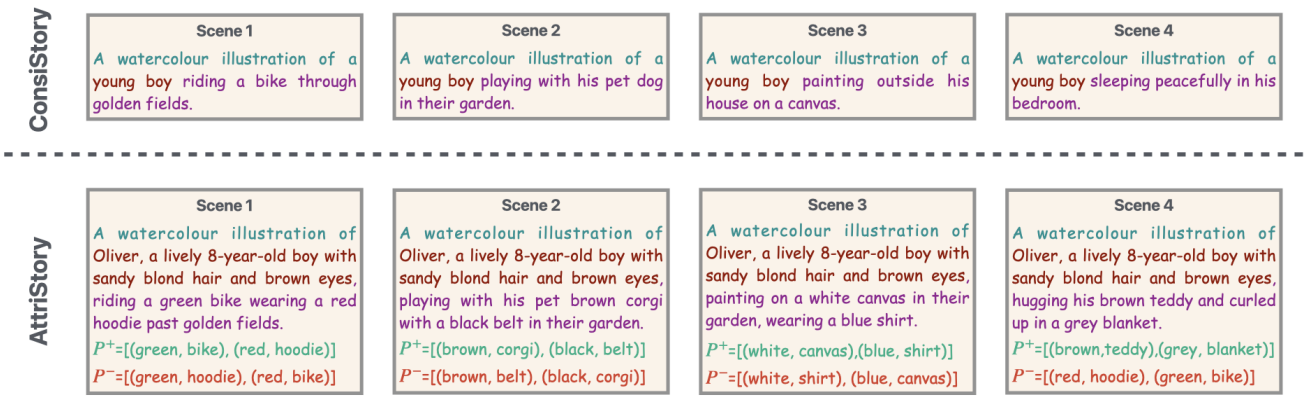

Figure 2. **Comparison of story narratives proposed in prior benchmarks vs. ours.** Existing approaches like ConsiStory (top) provide minimal visual specifications, capturing only basic character identity and actions. AttriStory (bottom) enriches narratives with explicit positive and negative attribute-object pairs (P^+ and P^-) for each scene, enabling systematic evaluation of fine-grained attribute realization. Oliver's story demonstrates how attributes like clothing color and accessories are specified and must be preserved across scenes.

lighting that cross-attention is the primary locus for identity information. IP-Adapter [39] introduces image-based prompting through decoupled cross-attention, maintaining compatibility with other controllable generation tools. The Chosen One [3] uses iterative identity clustering to identify images with similar appearance from a set of images generated by identical prompts, extracting a consistent character representation in an automated, prompt-guided manner. PhotoMaker [14] stacks identity embeddings to preserve identity while generalizing to unseen contexts, offering faster inference than test-time fine-tuning. However, these approaches require per-subject training, which is computationally expensive, hence limiting their scalability.

Recent works [18, 32, 40] has shifted focus towards training-free consistency mechanisms that are more practical for real-time storytelling applications. ConsiStory [32] introduces Subject-Driven Self-Attention (SDSA), which enforces character consistency by extending self-attention computations across all scenes. StoryDiffusion [40] proposes visual memory modules that capture long-range contextual information to stabilize both character appearance and scene progression. 1Prompt1Story [18] concatenates all story prompts into a single sentence with Singular-Value Reweighting (SVR) to maintain consistency, although the token length of text encoder restricts its expressiveness.

These methods represent substantial progress in preserving character identity across scenes. However, fine-grained attribute realization which is a critical need for professional storytelling, remains largely unaddressed. Our work bridges this gap by introducing both a comprehensive evaluation benchmark and a plug-and-play optimization method specifically designed for attribute realization, positioned as complementary to existing consistency approaches.

3. The AttriStory Benchmark

3.1. Motivation

Recent visual storytelling methods have made impressive progress in maintaining character consistency across scenes. Through attention mechanisms [18, 32, 40], these approaches successfully preserve character identity and appearance across narrative sequences. However, we've identified a critical gap: existing methods prioritize character consistency while largely neglecting fine-grained visual attributes specified in the narrative.

Current visual storytelling benchmarks [18, 32] typically capture narratives through simple, minimal descriptions, such as *"A photo of a young boy riding a bike through golden fields."* This level of specification is insufficient to capture how a storyboard artist would actually brief a visual team. In practice, artists communicate with rich and detailed specifications as *"A watercolor illustration of Oliver, a lively 8-year-old boy with sandy blond hair and brown eyes, riding a green bike wearing a red hoodie past golden fields."* These details convey the character personality and the relationship between attributes and objects in the scene. We illustrate the difference between ConsiStory (coarse-grained) and proposed AttriStory (fine-grained) story narratives in Figure 2, as existing benchmarks do not intend to capture this level of visual specificity.

We identify this as the *attribute realization problem*: ensuring that fine-grained attributes specified in the narrative text are faithfully rendered in generated images. This is fundamentally orthogonal to character consistency. While consistency mechanisms solve the problem of *the same character appearing in each scene*, attribute realization solves the problem of *how that character is accessorized*, the specific visual details that carry narrative significance.

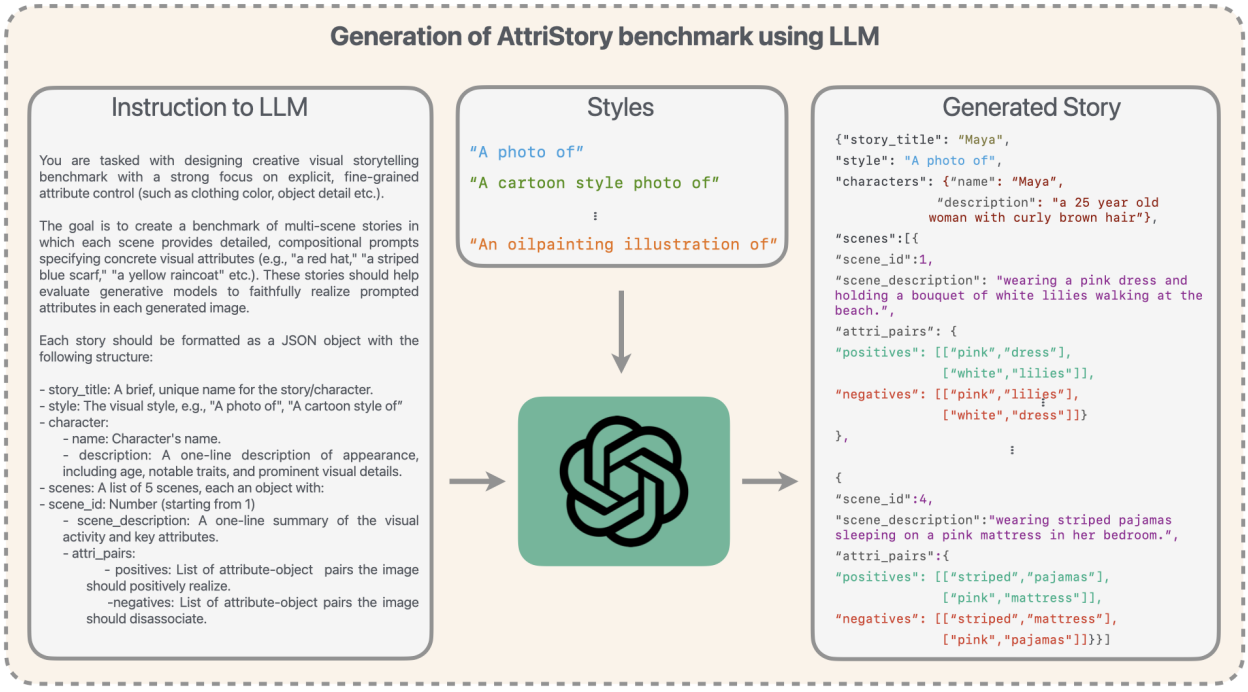

Figure 3. **LLM-driven benchmark generation.** The pipeline inputs artistic styles and structured instructions that emphasize explicit, fine-grained attribute specifications. For each story, the LLM chooses an artistic style and generates character descriptions, scene narratives, and positive (P^+) and negative (P^-) attribute-object pairs, producing structured stories enabling attribute realization.

To address this gap, we introduce *AttriStory*, a benchmark specifically designed to emphasize attribute realization in visual storytelling. This benchmark enables design of visual storytelling methods to systematically optimize for both character consistency and fine-grained visual adherence.

3.2. LLM-Driven Benchmark Generation

To scale benchmark creation while ensuring narrative coherence and realistic attribute specifications, we employ a large language model with structured prompts. We provide the LLM with detailed instructions specifying the story structure, artistic style, character description, and desired fine-grained attributes. As illustrated in Figure 3, we instruct the LLM to produce stories in the following structure:

1. **Character Description:** A detailed multi-attribute character specification (e.g., *"Maya, a 25-year-old boy with curly brown hair"*)
2. **Scene Narratives:** Scene-specific text descriptions enriched with fine-grained visual attributes.
3. **Positive Attribute-Object Pairs (P^+):** For each scene, attributes and objects that should co-occur (e.g., *["pink", "dress"], ["white", "lilies"]*). The attributes cover categories including colors, textures, materials etc.
4. **Negative Attribute-Object Pairs (P^-):** Negative pairs (e.g., *["pink", "lilies"], ["white", "dress"]*) to prevent models from spurious associations while generation.

We use ChatGPT [21] to generate the stories as illustrated in Figure 3. For generation, each scene prompt in a story is constructed in the following unified template: *"A [artistic style] of [character description], [scene description with fine-grained attributes]."* The resulting stories simulate realistic creative workflows, with the goal of adhering to fine-grained specifications made in the scene prompt, given detailed information as illustrated in Figure 2.

3.3. Dataset Statistics and Validation

We construct AttriStory benchmark comprising 200 multi-scene stories across 10 distinct artistic styles, each story containing 5 scenes. Each scene is annotated with 2 to 5 attribute-object pairs. The artistic styles include *photo, cartoon style, 3D animation, watercolor illustration, oil painting, crayon drawing, neon punk style, pixar-style, hyperrealistic digital painting, pastel color painting*.

To ensure high-quality story descriptions, we manually review the generated stories to ensure (1) the stories cover diverse attribute specifications in terms of color, texture, objects, (2) correctness of the attribute-object pairs, and (3) consistency between character descriptions across scene narratives. Stories failing these checks are manually revised to ensure they comply with the structure of the benchmark. This iterative process ensures that AttriStory provides a reliable benchmark to create rich visual story narratives.

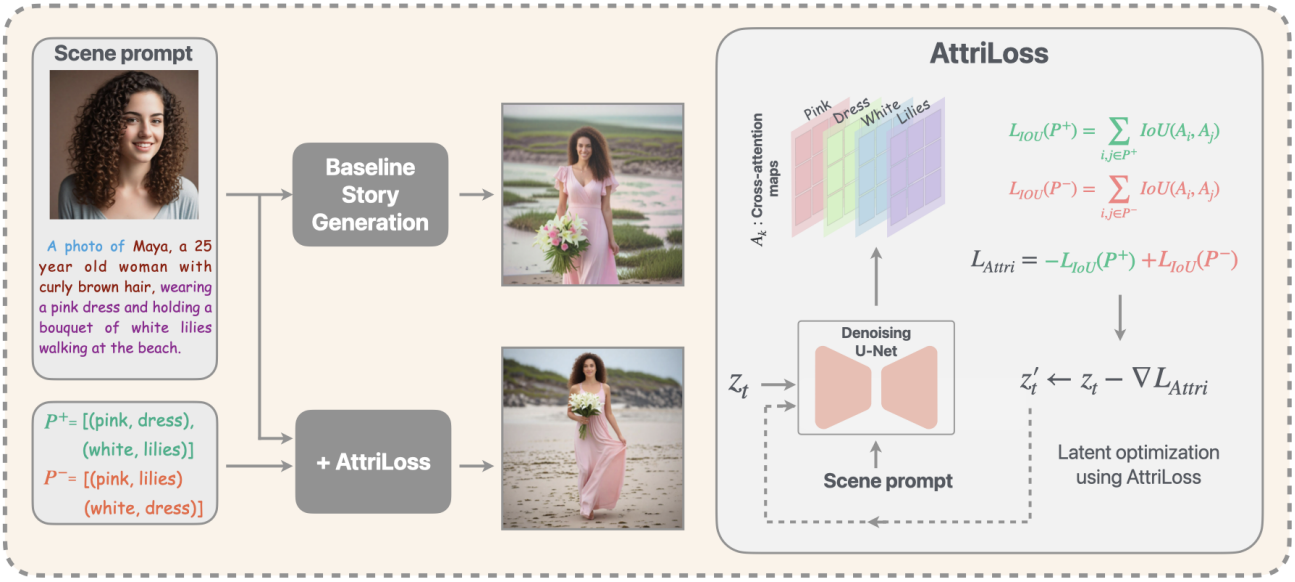

Figure 4. **AttriLoss: Targeted IoU loss on cross-attention maps.** Our method optimizes spatial overlap between attention maps of attribute-object token pairs during early denoising steps. By maximizing IoU for positive pairs (e.g., pink and dress should co-occur) and minimizing IoU for negative pairs (e.g., pink and lilies should not overlap), we guide the model to correctly localize fine-grained attributes.

4. AttriLoss: Targeted IoU loss for Attribute-Object Grounding in Visual Storytelling

Intuition. A key challenge in visual storytelling with diffusion models is imperfect alignment between prompt-specified attributes and generated visual objects. This typically arises due to the model's cross-attention mechanism: for a scene prompt like *"wearing a pink dress holding a bouquet of white lilies,"* attention maps for tokens such as *"pink"* and *"lilies"* may overlap spatially even though *"pink"* should only attend to *"dress"* and *"white"* to *"lilies."* Our key observation is that the cross-attention maps corresponding to tokens provide a direct visualization of attribute-object associations learned by the model. Any ambiguous overlaps in cross-attention maps of incorrectly associated tokens can result in images that fail to ground fine-grained attributes as specified. In Figure 5, the overlap of *"pink"* and *"lilies"* tokens result in creating pink roses which in undesired. By explicitly measuring and manipulating the spatial overlap between attention maps of attribute-object token pairs, we can guide the diffusion process to encourage correct and discourage spurious associations.

AttriLoss Formulation. In text-conditioned diffusion models, the cross-attention layers modulate image generation, providing implicit grounding signals linking text tokens to visual features. Each token in the prompt produces a spatial map indicating the degree of attention paid to different image regions during denoising. We construct an objective leveraging cross-attention maps to steer generation.

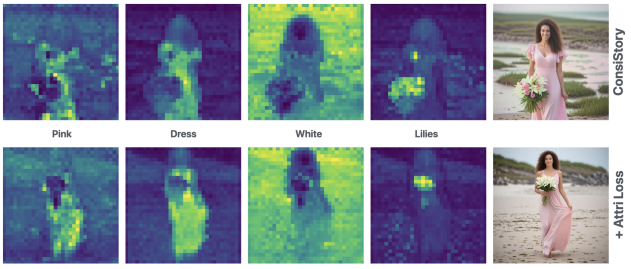

Figure 5. **Attention maps of ConsiStory and with AttriLoss.** The attention maps of baseline method ConsiStory show ambiguous spatial overlaps where attribute tokens *pink* and *lilies* attend to the same regions resulting in the image with pink roses as well. Using AttriLoss objective with ConsiStory, the attention maps for attribute-object pairs sharpen into distinct regions (*pink* and *lilies* don't overlap), achieving correct spatial localization.

Let P^+ be the set of positive attribute-object token pairs that should co-occur spatially (e.g., (*"pink"*, *"dress"*)), and P^- be the set of negative pairs that should not spatially overlap (e.g., (*"pink"*, *"lilies"*)). For each text token k, let $A_k \in \mathbb{R}^{H \times W}$ be its spatial cross-attention map aggregated across U-Net layers while denoising.

The Intersection-over-Union (IoU) between two attention maps A_i and A_j is calculated as:

$$\text{IoU}(A_i, A_j) = \frac{A_i \odot A_j}{\max(A_i + A_j - A_i \odot A_j, \epsilon)} \quad (1)$$

where $\odot$ denotes element-wise multiplication and ϵ is a small constant for numerical stability.

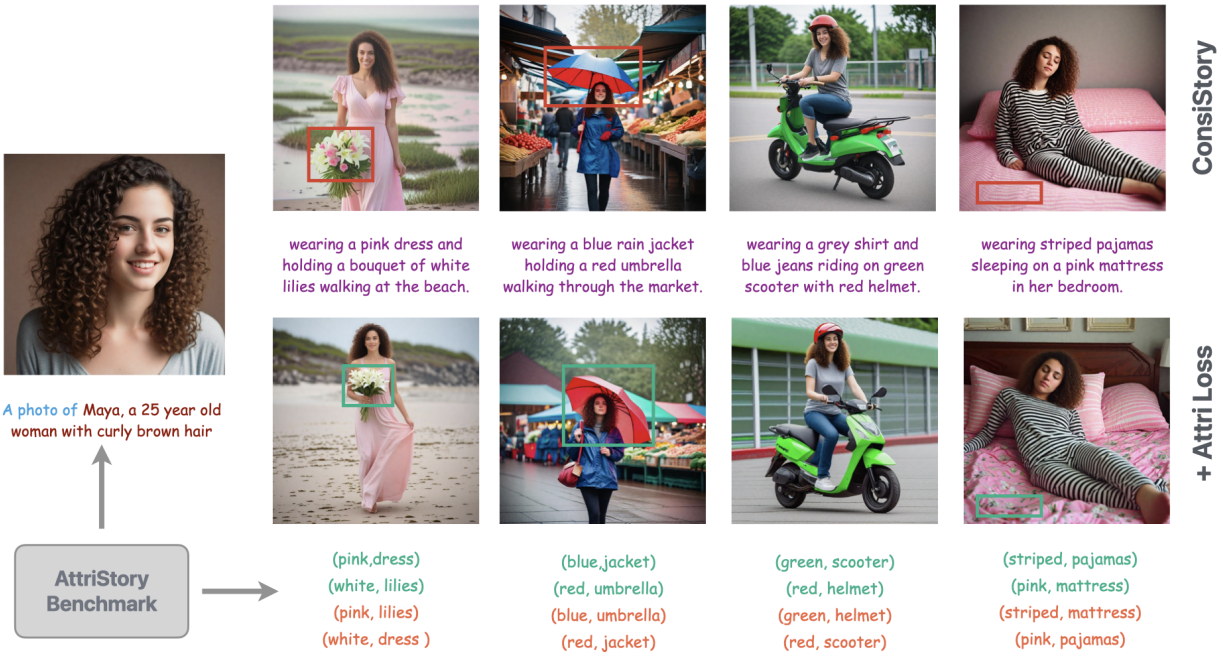

Figure 6. **Qualitative results of ConsiStory baseline with and without AttriLoss.** Using ConsiStory, the character consistency is maintained but it fails to correctly bind fine-grained attributes (e.g., pink roses are rendered with white lilies (1), umbrella is partially colored as blue instead of red (2)). With AttriLoss, attribute specifications are faithfully realized while preserving character consistency.

The AttriLoss objective is formulated as:

$$\mathcal{L}_{\text{Attri}} = -\sum_{(i,j)\in P^+} \text{IoU}(A_i, A_j) + \sum_{(i,j)\in P^-} \text{IoU}(A_i, A_j) \tag{2}$$

The latent code $\mathbf{z}_t$ is updated via gradient descent through the proposed $\mathcal{L}_{\text{Attri}}$ objective as:

$$\mathbf{z}_t^{'} \leftarrow \mathbf{z}_t - \nabla_{\mathbf{z}_t}\mathcal{L}_{\text{Attri}} \tag{3}$$

This latent optimization steers the denoising process to maximize the IoU of positive attribute-object pairs P^+ and minimize the IoU of negative attribute-object pairs P^-. We perform this update in the early denoising timesteps (timesteps 1 to 25 out of 50 total steps) which are critical for establishing structure and semantic content [8], enabling grounding of object colors and textures as specified.

Integration with Existing Pipelines Our AttriLoss objective is designed as a plug-and-play module compatible with existing storytelling pipelines that leverage self-attention mechanisms to enforce consistency in training-free manner. We integrate our module with Vanilla SDXL [24], ConsiStory [32] and StoryDiffusion [40]. It operates independently for each scene, refining latent code while denoising to improve attribute adherence while preserving the consistency enforced by the baseline storytelling methods. This orthogonal design allows immediate adoption without requiring architectural modifications or retraining, enabling straightforward integration into existing workflows.

5. Experiments and Results

5.1. Implementation Details

We evaluate three primary baselines: (1) Vanilla SDXL without consistency mechanisms, (2) StoryDiffusion with visual memory for cross-scene consistency, and (3) ConsiStory with Subject-Driven Self-Attention. We report the performance of each baseline and on integrating AttriLoss with it (denoted as "+ AttriLoss") in Table 1. Although, we evaluate the performance of 1Prompt1story [18] on our benchmark, we observe that its performance is limited in stories with detailed descriptions as ours, due its token length constraint. All experiments are conducted using Stable Diffusion XL (SDXL) as the generative model, following [32, 40]. AttriLoss optimizes latent codes during early denoising upto 25 time steps, using AdamW optimizer with 0.01 learning rate. All experiments are done on a single NVIDIA A6000 GPU. Evaluation is performed on the AttriStory benchmark (Section 3) with 200 multi-scene stories across 10 artistic styles generated using ChatGPT [21].

5.2. Evaluation Metrics

Visual storytelling evaluation must be assessed in multiple perspectives: fine-grained attribute realization, cross-scene consistency, and overall visual quality. Here, we describe the four complementary metrics we employ, which provide a comprehensive evaluation of visual storytelling methods.

Figure 7. **Qualitative results of StoryDiffusion baseline with and without AttriLoss.** Using Consistory (top), the character consistency is maintained but fails to correctly bind fine-grained attributes (e.g., grey coat (2), yellow coat(3) and beige jacket(3) are not rendered using StoryDiffusion). With AttriLoss (bottom), attribute specifications are faithfully realized while character consistency is preserved.

Image-Text Alignment. We use two metrics to measure whether generated images adhere to textual scene descriptions. CLIP Image-Text Similarity (CLIP-T) [9] computes cosine similarity between image and text embeddings from CLIP, providing a standard baseline. However, embedding-based metrics compute similarity without specific focus on compositional details, limiting their effectiveness for fine-grained attribute prompts. VQAScore [15] reformulates alignment as visual question answering task: given an image and text, it converts descriptions into questions (e.g., *"Is the dress pink?"*) and measures the probability of a "Yes" response from a VQA model. This is better aligned with attribute-centric prompts proposed in our benchmark.

Cross-Scene Consistency. CLIP Image-Image Similarity (CLIP-I) measures pairwise visual similarity between all generated images within a story, capturing whether character appearance and visual features persist across scenes, independent of text alignment.

Perceptual Quality. DreamSim [5] provides perceptual similarity aligned with human judgments, capturing mid-level visual properties including layout, object pose, color, and attribute variations. Unlike pixel-level metrics, DreamSim reflects holistic narrative quality.

Together, these metrics enable rigorous evaluation across attribute realization (VQAScore, CLIP-T), character consistency (CLIP-I), and quality measure (DreamSim).

Method	VQA-Score↑	CLIP-T↑	CLIP-I↑	DreamSim↑
1Prompt1Story [18]	0.8117	0.3816	0.8410	0.6929
Vanilla SDXL [24]	0.7957	0.3696	0.8188	0.6760
+ AttriLoss	**0.8225**	**0.3775**	**0.8517**	**0.7170**
StoryDiffusion [40]	0.8363	**0.3912**	0.8301	0.6925
+ AttriLoss	**0.8636**	0.3874	**0.8553**	**0.7215**
ConsiStory [32]	0.8136	0.3871	0.8494	0.7326
+ AttriLoss	**0.8490**	**0.3909**	**0.8667**	**0.7555**

Table 1. **Quantitative comparison** of evaluation metrics on integrating AttriLoss with prior visual storytelling methods.

5.3. Quantitative Results

Table 1 presents comprehensive quantitative comparisons across all baselines and metrics. AttriLoss consistently improves baseline performance. Significant improvement in VQAScore is associated with better fine-grained attribute realization, as it directly measures whether the specified attributes bind correctly to objects in generated images. CLIP-T shows similar or modest gains, which is expected since it is a coarse metric that does not capture compositional details, indicating that AttriLoss preserves global semantic alignment while refining attribute-level specificity. Importantly, CLIP-I scores remain strong, confirming that attribute grounding does not compromise cross-scene character consistency and validating our design that attribute realization operates orthogonally to existing consistency

Figure 8. **Attribute realization across diverse stories** using baseline as ConsiStory (top) and with AttriLoss (bottom). Each column shows a scene in varied artistic styles (Pixar, cartoon, oil painting, photo, watercolor). AttriLoss corrects attribute-object binding failures: peacock's red velvet capelet (1), Dr. Barkley's glasses (2), yellow flag on the raft (3), Luke's green hoodie (4), Oliver's green bike (5)

mechanisms. DreamSim improves across all methods, indicating that AttriLoss enhances both attribute-specific alignment and overall perceptual quality. We observe that *ConsiStory + AttriLoss* achieves the best overall performance by balancing consistency and improved attribute realization.

5.4. Qualitative Analysis

Figure 6 demonstrates ConsiStory with and without AttriLoss on a full multi-scene narrative. The baseline preserves character consistency but fails to bind fine-grained attributes correctly:*"white lilies"* are incorrectly rendered as *"pink roses and white lilies"* in scene 1 and *"red umbrella"* appears incorrectly colored as *"red and blue"* in scene 2. With AttriLoss, attribute-object associations are properly grounded while character consistency is maintained across all scenes. Similarly, Figure 7 shows StoryDiffusion with and without AttriLoss, focusing on attribute realization. Without AttriLoss, the attributes are not realized in scenes 2 to 4. With AttriLoss, compositional attributes are rendered (*"grey coat", "yellow coat", "beige jacket"*) and consistency is maintained across scenes.

Beyond full-narrative comparisons, in Figure 8, we presents selected scenes from different stories spanning varied visual styles: pixar illustration, cartoon, oil painting, photo, and watercolor. In each case, the baseline generates plausible characters and scenes but fails to correctly realize key attribute-object pairs: *a peacock with a woman sitting on it instead of it wearing a red velvet capelet (1), Dr. Barkley's glasses are missing (2), the otter's boat missing its yellow flag (3), Luke hoodie is beige and green in color (4), Oliver's bike in red instead of green (5)*. With AttriLoss, these fine-grained attribute specifications are faithfully realized across styles, demonstrating its effectiveness.

Limitations. While attributes are better realized on integrating AttriLoss, there remain cases where the action described is not realized in the generated images. In Figure 8, *Dr.Barkley's face is front facing rather than reading a book (scene 2) and Otto, the otter is not actually steering his raft*, demonstrating further scope of improvement.

6. Conclusion

This work addresses a key gap in visual storytelling with diffusion models by focusing on fine-grained attribute realization alongside character consistency. We introduce AttriStory, a benchmark of 200 multi-scene stories with explicit fine-grained attribute-object annotations across 10 artistic styles, enabling attribute realization for visual storytelling in a systematic manner. Complementing this, we propose AttriLoss, a targeted IoU loss applied to cross-attention maps during early diffusion steps that guides models to accurately localize specified attributes. Extensive experiments demonstrate significant improvements across all baselines. By bridging identity preservation and attribute control, this work advances high-fidelity visual storytelling.

References

[1] Kiymet Akdemir and Pinar Yanardag. Oracle: Leveraging mutual information for consistent character generation with loras in diffusion models. *arXiv preprint arXiv:2406.02820*, 2024. 2

[2] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 2

[3] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, pages 1–12, 2024. 3

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2

[5] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 7

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[7] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, et al. Interactive story visualization with multiple characters. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 1

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 6

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. 7

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2

[12] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 2

[13] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[14] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8640–8650, 2024. 3

[15] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 7

[16] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7817–7826, 2024. 2

[17] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6190–6200, 2024. 1

[18] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-to-image generation using a single prompt. *arXiv preprint arXiv:2501.13554*, 2025. 1, 2, 3, 6, 7

[19] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European conference on computer vision*, pages 70–87. Springer, 2022. 1

[20] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 1

[21] OpenAI. Chatgpt. https://chatgpt.com/, 2025. Large language model. 4, 6

[22] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2920–2930, 2024. 1

[23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6, 7

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image gener-

ation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6527–6536, 2024. 2

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[32] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 43(4):1–18, 2024. 1, 2, 3, 6, 7

[33] Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. Characonsist: Fine-grained consistent character generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16058–16067, 2025. 1

[34] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[35] Qinghe Wang, Baolu Li, Xiaomin Li, Bing Cao, Liqian Ma, Huchuan Lu, and Xu Jia. Characterfactory: Sampling consistent characters with gans for diffusion models. *IEEE Transactions on Image Processing*, 2025. 2

[36] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848*, 2023. 2

[37] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2

[38] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1850–1860, 2025. 1

[39] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 3

[40] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37: 110315–110340, 2024. 1, 2, 3, 6, 7

[41] Zhengguang Zhou, Jing Li, Huaxia Li, Nemo Chen, and Xu Tang. Storymaker: Towards holistic consistent characters in text-to-image generation. *arXiv preprint arXiv:2409.12576*, 2024. 2