



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

Manogna Sreenivas; Rohit Kumar; Soma Biswas
Indian Institute of Science, Bengaluru, India

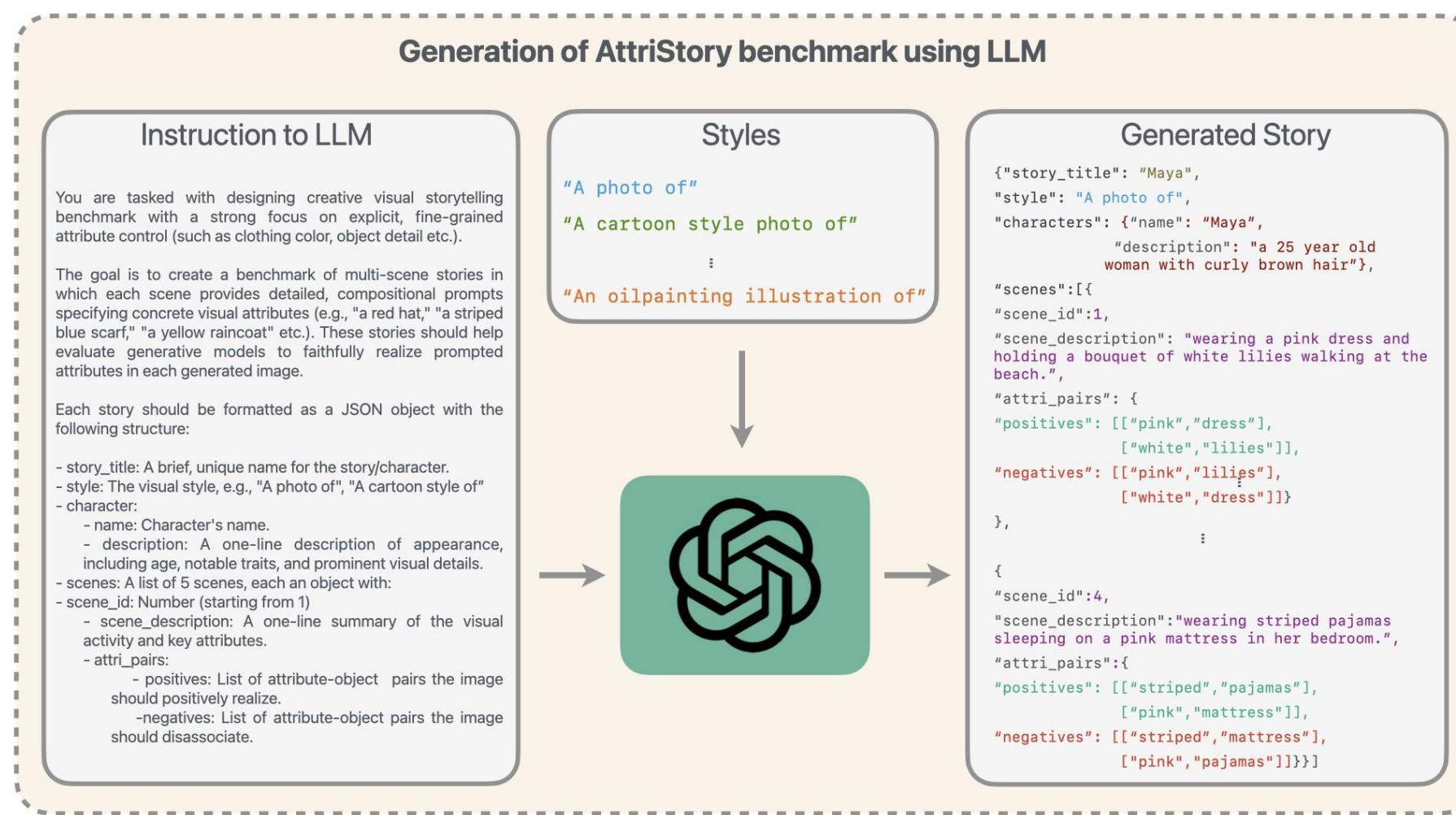
AttriStory

- Diffusion-based visual storytelling methods achieve strong character consistency across scenes, yet fine-grained attributes remain systematically unrealized and unevaluated.
- We introduce AttriStory benchmark and AttriLoss, a cross-attention IoU objective for fine-grained attribute realization in visual storytelling.



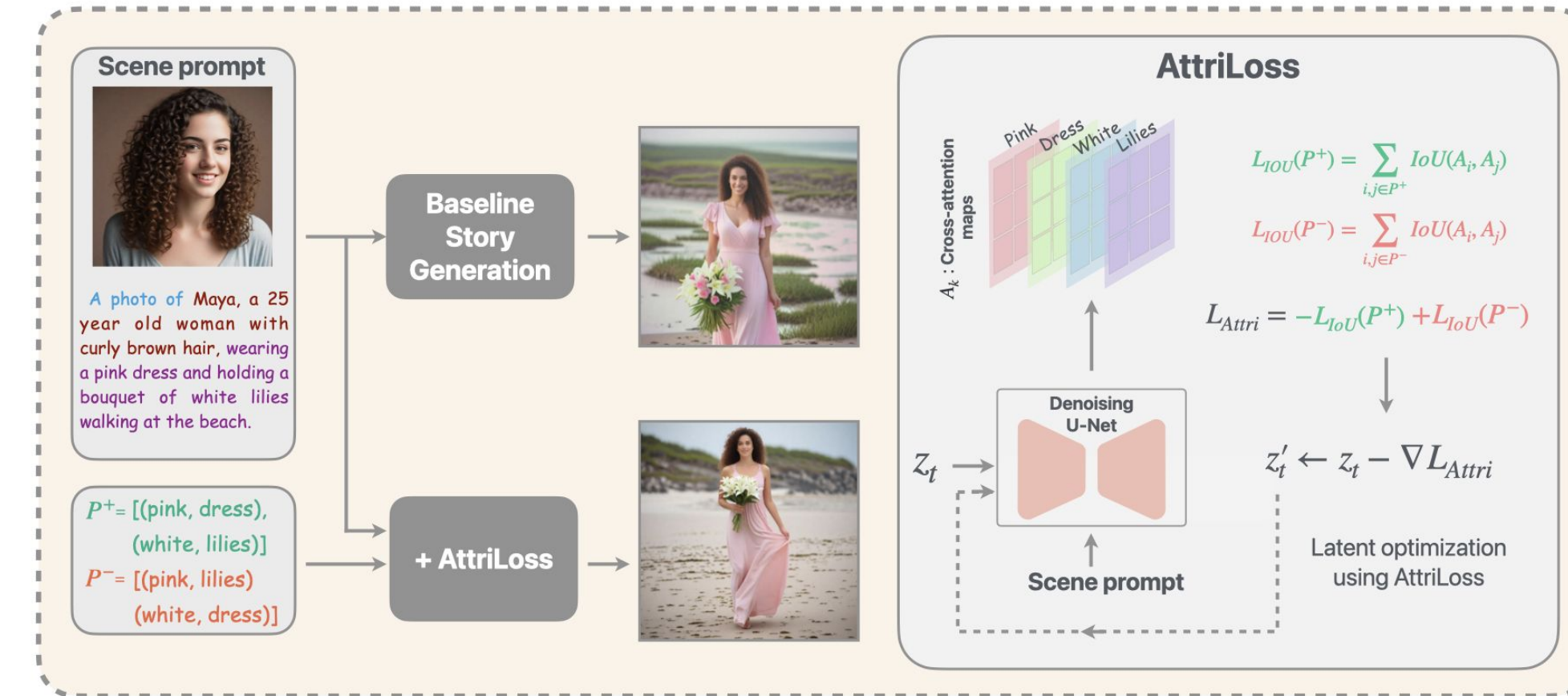
AttriStory benchmark: character consistency meets fine-grained attribute realization across scenes.

- Character Description:** A detailed multi-attribute character specification
- Scene Narratives:** Scene-specific text descriptions enriched with fine-grained visual attributes.
- Positive Attribute-Object Pairs (P+):** For each scene, attributes and objects that should co-occur. The attributes cover categories including colors, textures, materials etc.
- Negative Attribute-Object Pairs (P-):** Negative pairs to prevent models from spurious associations while generation.



LLM-driven benchmark generation of AttriStory

AttriLoss



AttriLoss optimizes cross-attention IoU for fine-grained attribute realization.

- IoU between two attention map:

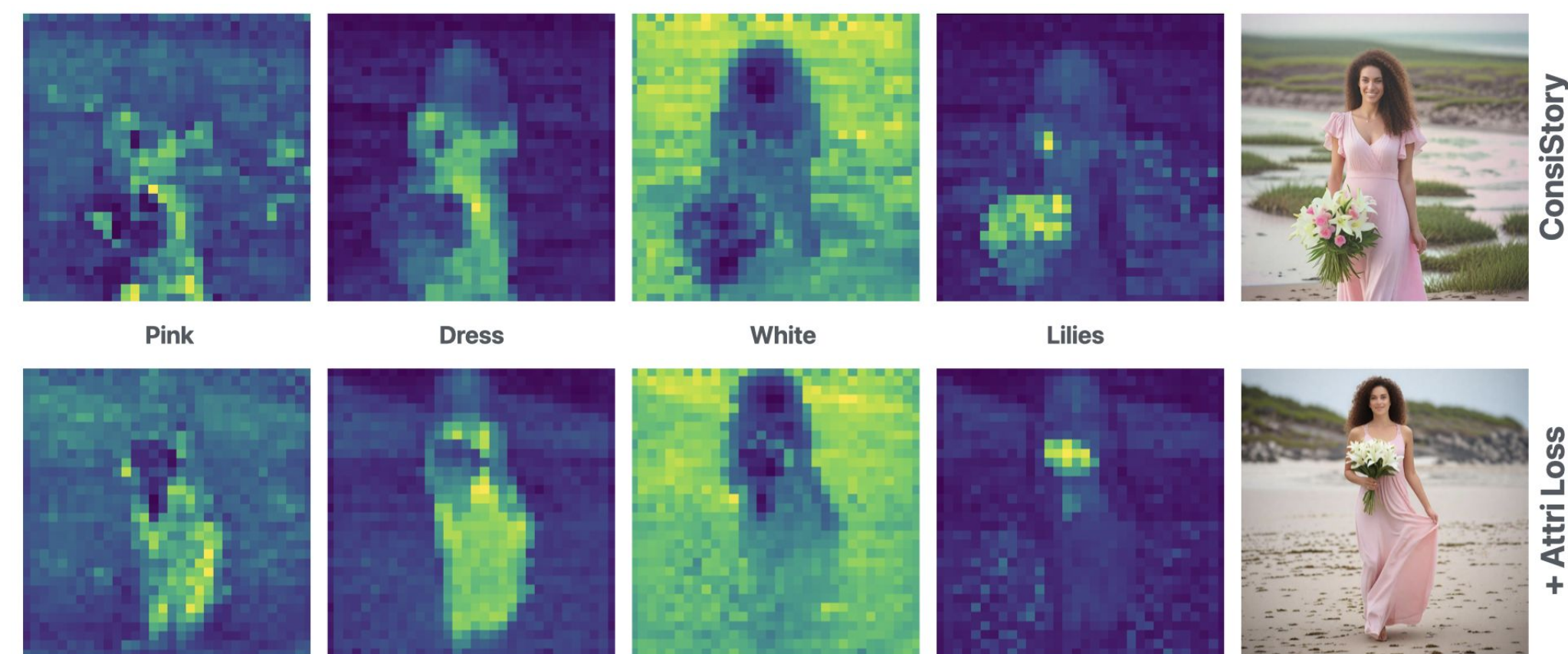
$$IoU(A_i, A_j) = \frac{A_i \odot A_j}{\max(A_i + A_j - A_i \odot A_j, \epsilon)}$$

- AttriLoss Objective:

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

- Latent Optimization:

$$z'_t \leftarrow z_t - \nabla_{z_t} \mathcal{L}_{Attri}$$



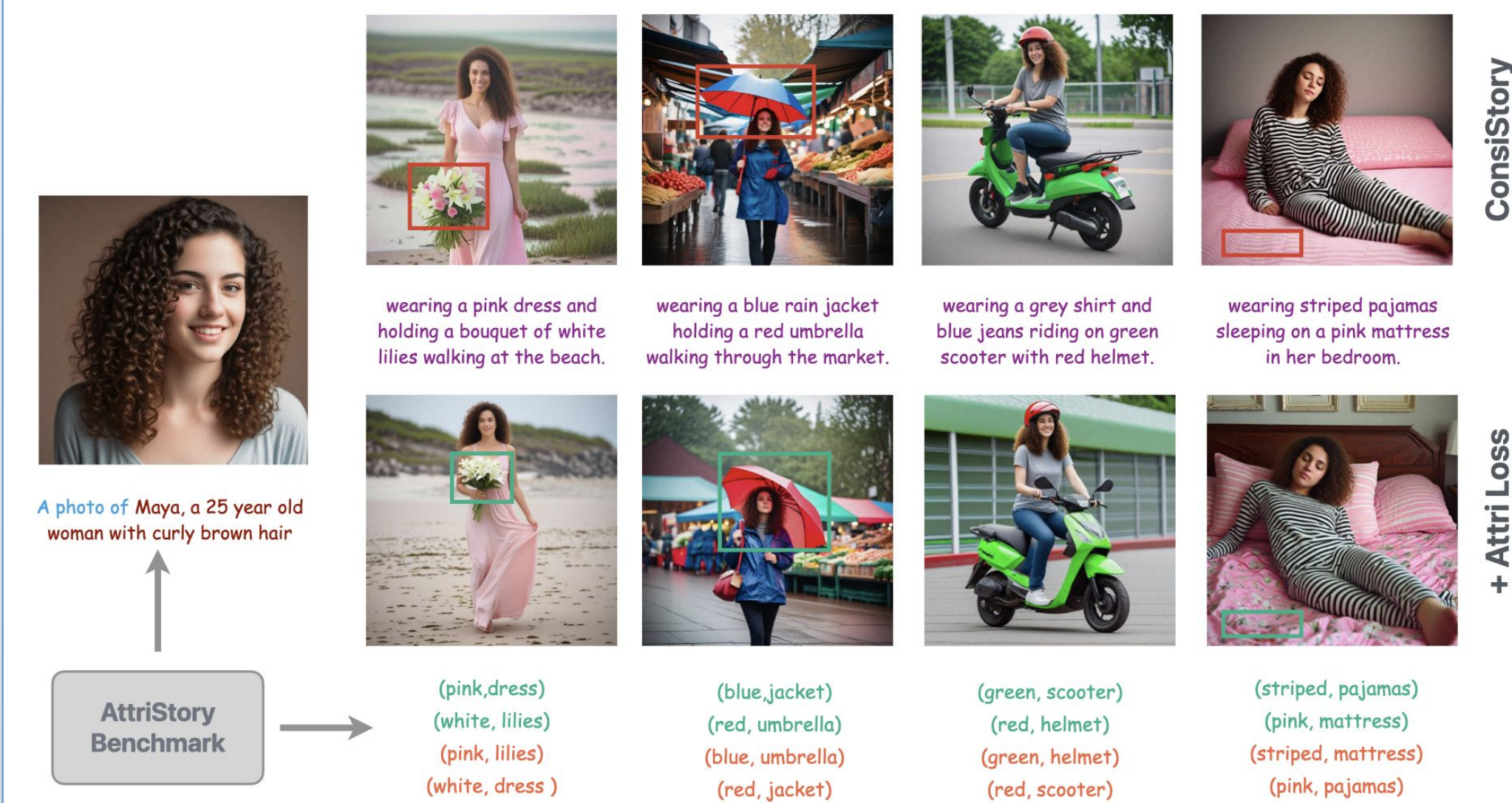
- AttriLoss is plug-and-play method to enhance visual storytelling without the need for retraining or architectural changes.

Results

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

Quantitative comparison of AttriLoss integrated with prior visual storytelling methods.

- VQAScore & CLIP-T measure attribute realization, CLIP-I measures character consistency, and DreamSim measures image quality.
- AttriLoss consistently improves VQA-Score, CLIP-T and DreamSim across all baselines, confirming better attribute realization and perceptual quality
- CLIP-I remains strong, confirming attribute realization does not compromise character consistency.

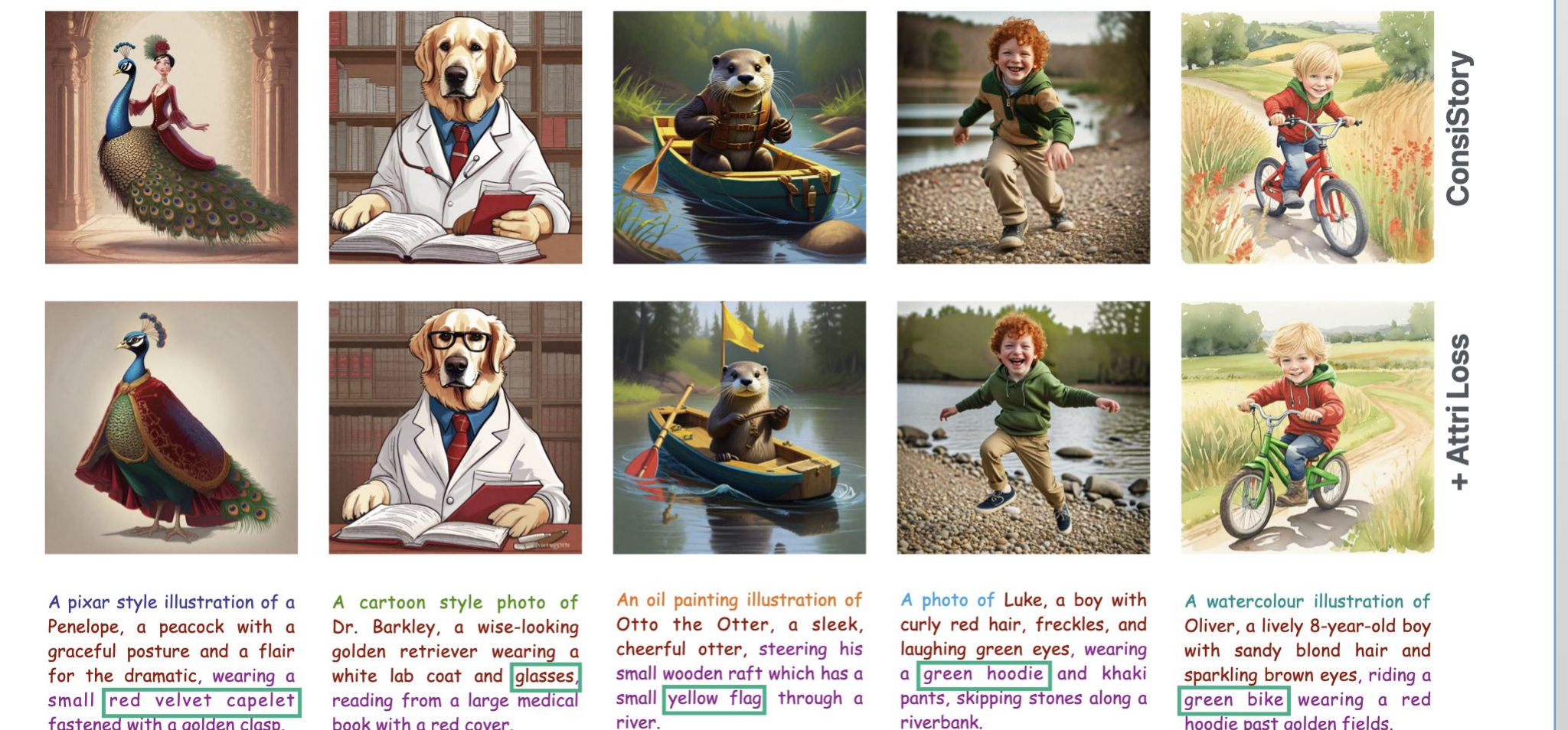


Qualitative results of Consistency baseline with and without AttriLoss.

Results



Qualitative results of StoryDiffusion baseline with and without AttriLoss.



Attribute realization across diverse stories.

Conclusions

- AttriStory benchmarks fine-grained attribute realization across 200 multi-scene stories and 10 artistic styles
- AttriLoss guides accurate attribute localization via cross-attention IoU during early diffusion steps
- Consistent improvements across all baselines, bridging identity preservation and attribute control
- Limitation: Actions described in the prompt might still fail to be realized and this remains a scope for future work.