RealDiffusion: Physics-informed Attention for Multi-character Storybook Generation

Supplementary Material

Overview

In this supplementary material, we provide extensive additional experimental results to further substantiate the effectiveness and robustness of **RealDiffusion**. The document is organized as follows:

- **Appendix A** presents additional qualitative comparisons against state-of-the-art methods. These visual results further validate our framework's superiority in maintaining character consistency while preserving narrative dynamism across diverse scenarios.
- **Appendix B** conducts a comprehensive ablation analysis detailing the effectiveness of insulated heat diffusion and region-aware stochasticity.
- **Appendix C** explores the robustness and generalization capabilities of our framework by demonstrating its stability across different random seeds, various community fine-tuned SDXL backbones, and distinct identity inputs.
- **Appendix D** showcases the potential of RealDiffusion in generating long-form narratives, proving its ability to sustain character consistency over extended sequences.

A. Additional Main Experiments

In this section, we present additional qualitative comparisons to further demonstrate the superiority of RealDiffusion over existing state-of-the-art approaches. We compare our method against seven leading baselines: IP-Adapter [8], PhotoMaker [3], StoryDiffusion [9], ConsiStory [6], One-Prompt-One-Story (1P1S) [4], Zigzag Sampling [2], and CharaConsist [7]. To comprehensively evaluate the performance, we selected six diverse prompt sets covering both multi-character interactions and single-character narratives. The results regarding multi-character interactions are visualized in Figures 1 and 2, while single-character narratives are presented in Figure 3.

Multi-character Scenarios. As shown in Figures 1 and 2, the first two prompt sets involve complex interactions between two distinct characters. This setting poses a significant challenge for existing methods which frequently suffer from identity leakage or attribute swapping where the features of one character blend into the other. Baselines such as IP-Adapter and PhotoMaker tend to average these features leading to characters that appear indistinguishable. In contrast, RealDiffusion effectively utilizes dynamic subject masks and insulated heat diffusion to disentangle identities. This approach ensures that each character maintains its unique visual appearance while interacting naturally within the scene and successfully prevents visual traits from bleeding into neighboring regions.

Single-character Narratives. The subsequent prompt sets illustrated in Figure 3 focus on single-character generation across varying scenes and artistic styles. The primary challenge here lies in balancing robust identity preservation with the diversity of poses and backgrounds required by the prompts. While baseline methods often struggle to maintain narrative dynamism and frequently result in repetitive poses or rigid facial expressions, RealDiffusion addresses this balance effectively. Our framework preserves high-fidelity details including specific clothing textures, hairstyles, and facial features across frames. Furthermore, it allows the character to faithfully adhere to the text instructions for complex actions and scene changes, enabling the generation of diverse poses that strictly follow the narrative description.

B. Additional Ablation Studies

In this section, we provide a more detailed analysis of the individual contributions of each component within the RealDiffusion framework. We specifically examine the visual impact of the insulated heat diffusion mechanism and the region-aware stochasticity process. The qualitative comparisons illustrating the necessity of these core modules are presented in Figure 4.

B.1. Impact of Individual Modules

Effectiveness of insulated heat diffusion. To strictly validate the necessity of physical regularization, we analyze the generation results when the framework operates solely under the guidance of region-aware stochasticity. In this specific configuration, the system lacks the dissipative smoothing kernel inherently provided by the heat equation. Consequently, despite the presence of identity signals, we observe significant high-frequency artifacts and inconsistent attributes such as flickering clothing details or unstable facial structures between adjacent frames. These visual discrepancies confirm that the insulated heat diffusion mechanism is indispensable for suppressing temporal noise and ensuring a smooth and coherent visual progression throughout the entire narrative sequence.

Effectiveness of region-aware stochasticity. Conversely, we assess the performance when the system is constrained to operate using only insulated heat diffusion to regulate temporal consistency. While this configuration effectively ensures strong character coherence and minimizes identity drift, it simultaneously introduces a risk of over-smoothing

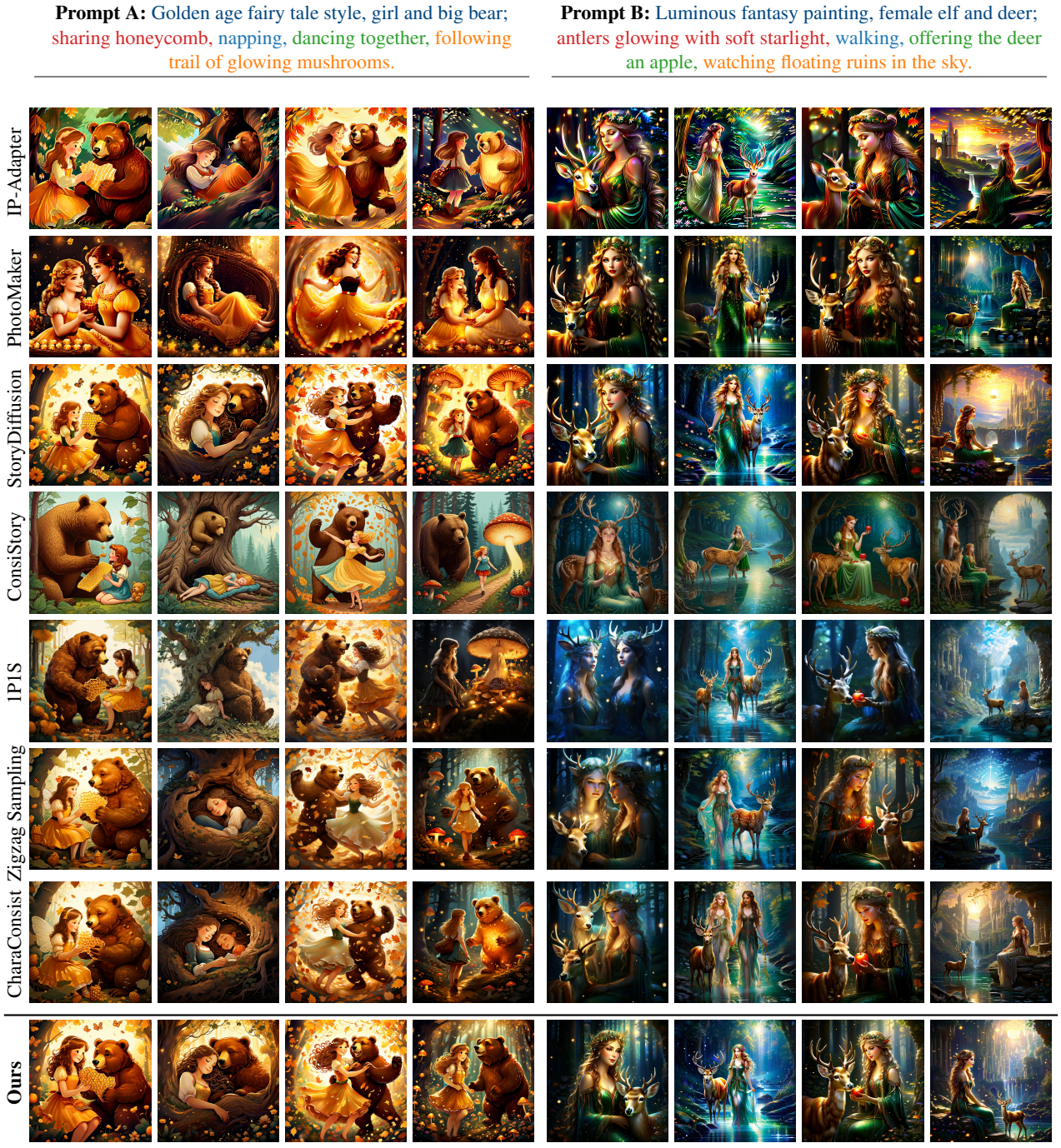

Figure 1. A qualitative comparison of our RealDiffusion against seven state-of-the-art baseline models in multi-character scenarios.

features which actively suppresses narrative motion. The resulting images tend to appear overly rigid or stagnant as the model fails to break away from the averaged feature mean without the complementary random perturbations. This limitation demonstrates that the region-aware stochasticity is essential for exploring the latent space to enable natural pose variations and dynamic background evolution as dictated by the text prompts.

C. Robustness and Generalization

In this section, we explore the robustness and generalization capabilities of the RealDiffusion framework. Our evaluation covers three key aspects: stability across varying ran-

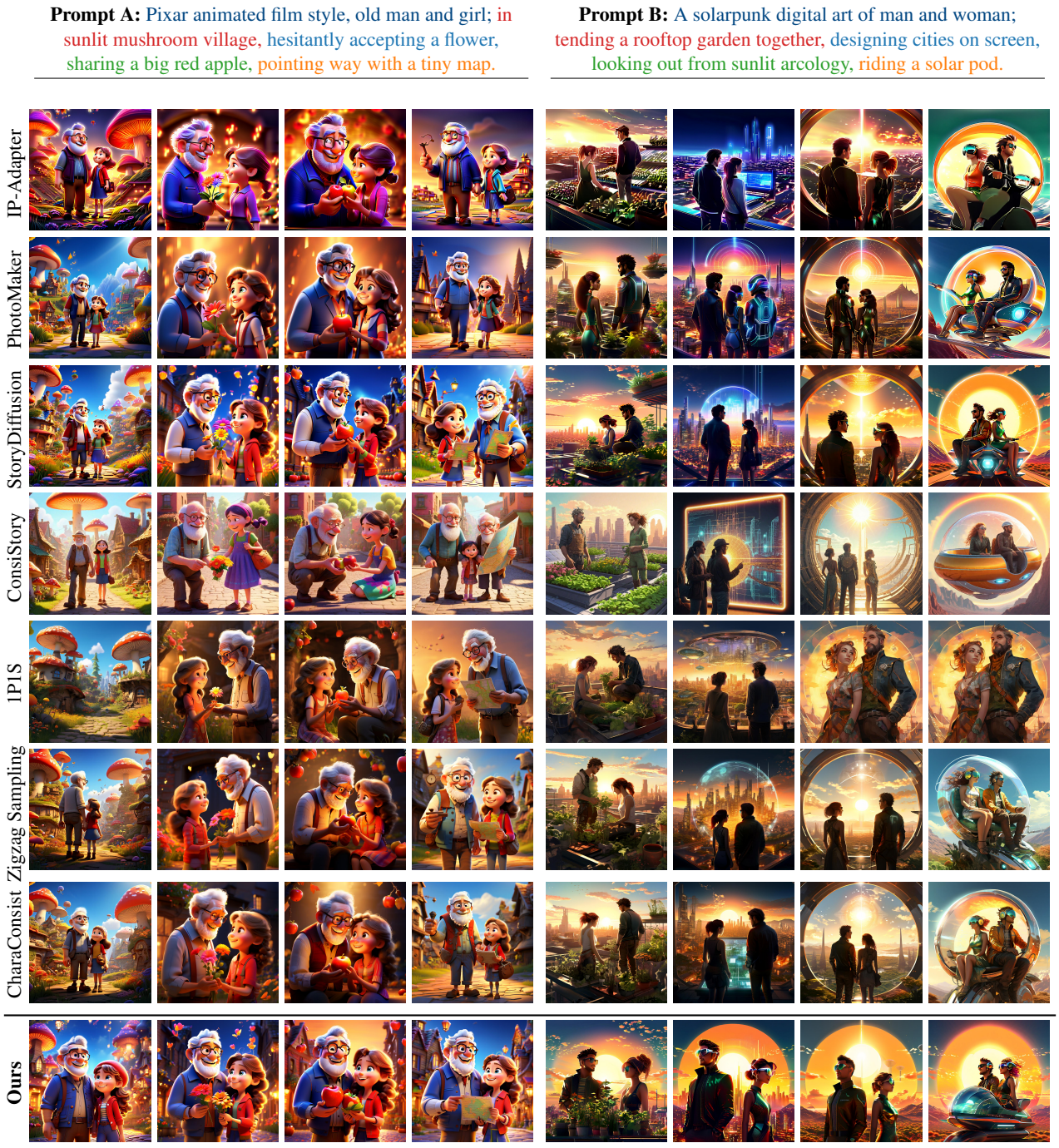

Figure 2. A qualitative comparison of our RealDiffusion against seven state-of-the-art baseline models in multi-character scenarios.

dom initialization seeds, compatibility with diverse community fine-tuned SDXL backbones, and the distinctiveness of character identities under identical narrative constraints.

C.1. Impact of Random Seeds

Diffusion models are inherently stochastic as the initial noise sampling significantly influences the final image layout and composition. To ensure that the consistency of our method is not an artifact of cherry-picked seeds, we tested the same prompt across multiple distinct random seeds.

As illustrated in Figure 5, while changing the seed alters the specific pose, background composition, and lighting details as expected in generative models, RealDiffusion consistently maintains the identity of the subject. The facial

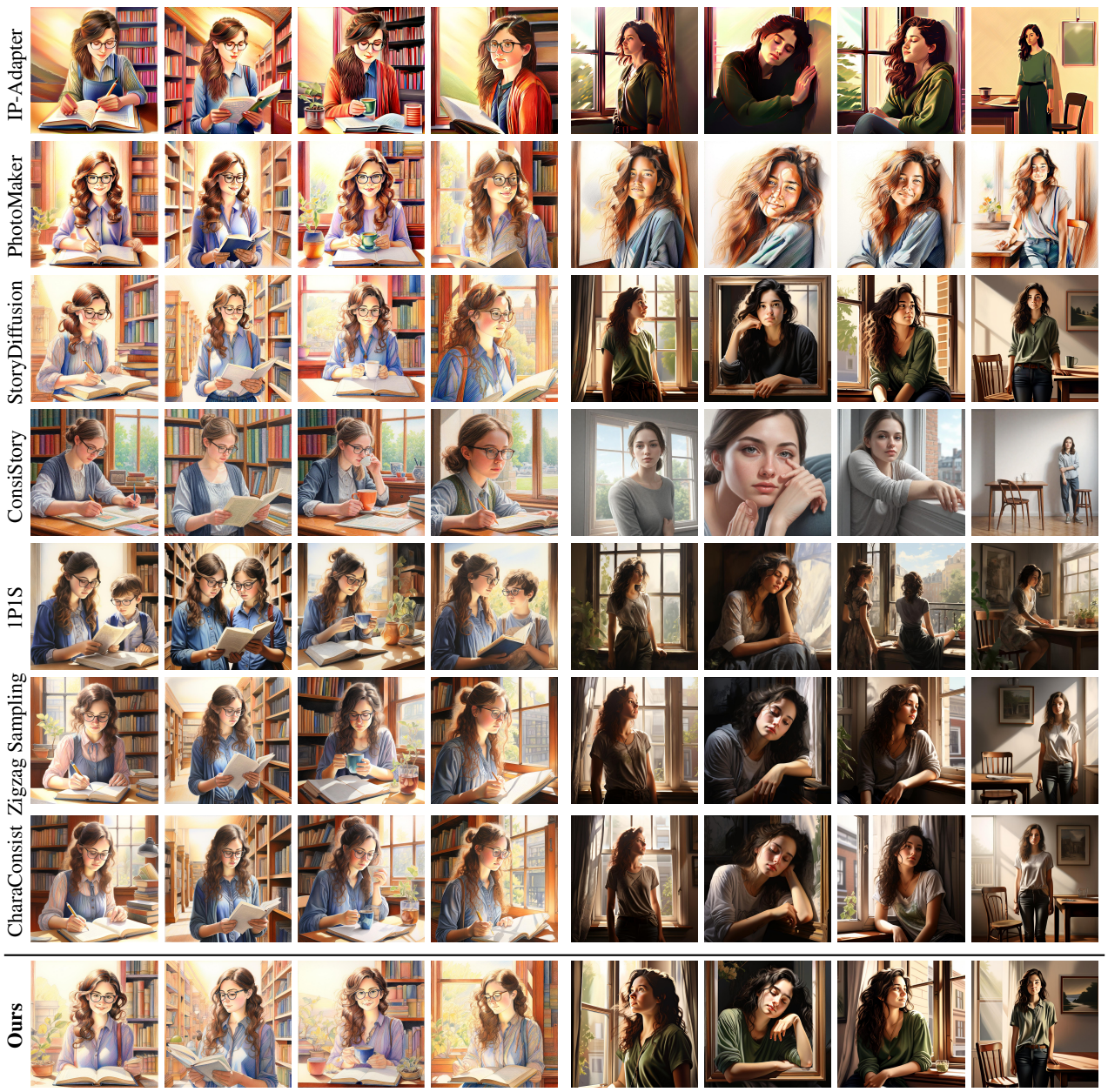

Figure 3. A qualitative comparison of our RealDiffusion against seven state-of-the-art baseline models in single-character.

features, clothing style, and overall character attributes remain robust across all varying initializations. This confirms that our Physics-informed Attention mechanism provides a stable regularization constraint that operates effectively regardless of the initial noise distribution.

C.2. Compatibility with Different Backbones

A key advantage of RealDiffusion is its training-free nature, which allows it to be seamlessly integrated into various community-finetuned models without the need for parameter fine-tuning. We evaluated our framework on the

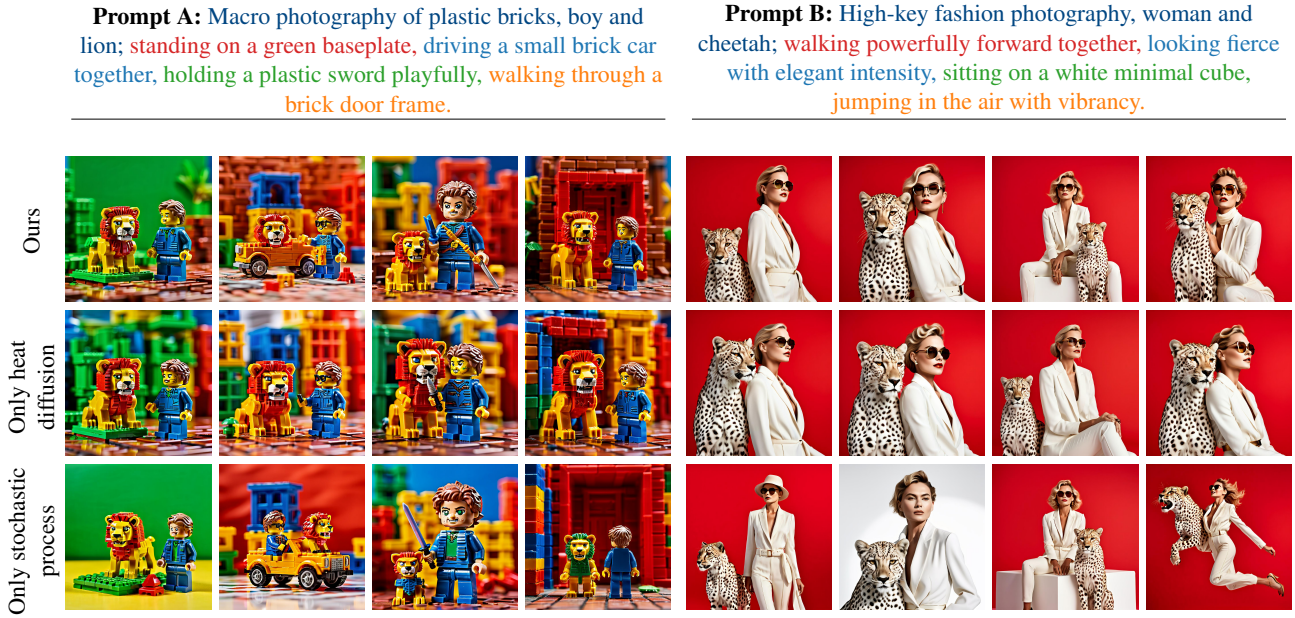

Figure 4. A qualitative ablation study demonstrating the effectiveness of insulated heat diffusion and region-aware stochasticity.

Figure 5. A qualitative evaluation of robustness across different random initialization seeds.

official Stable Diffusion XL [5] and three popular community checkpoints including Playground V2.5 [1], RealVisXL V4.0, and Juggernaut XL V9.

The qualitative results illustrating this compatibility are presented in Figure 6.

- **Playground V2.5**: Playground V2.5 is known for its high aesthetic quality and vibrant color palettes; our method preserves its artistic rendering style while effectively fixing the character identity.
- **RealVisXL V4.0**: RealVisXL V4.0 focuses on hyper-

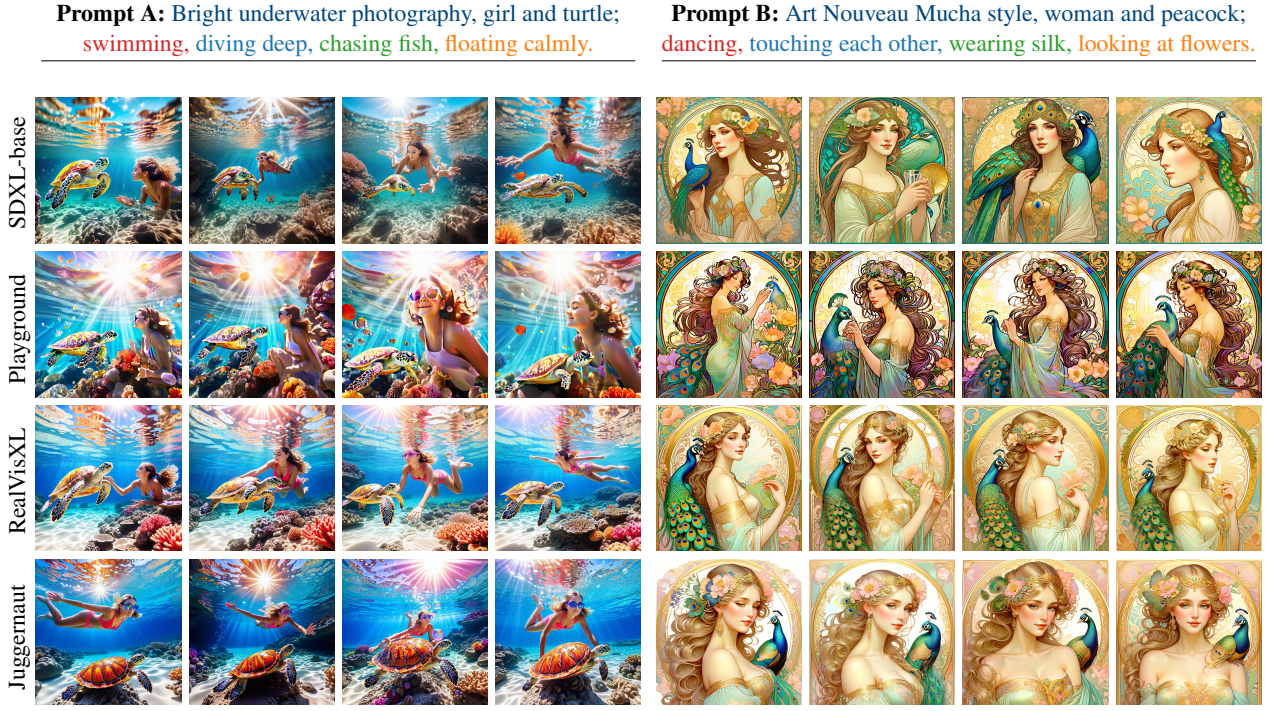

Figure 6. A qualitative evaluation of compatibility across different community fine-tuned SDXL backbone models.

realistic photography, and our method ensures that the intricate skin textures and realistic lighting effects on the subject remain consistent.

- **Juggernaut XL V9**: Juggernaut XL V9 is optimized for cinematic composition, and our method successfully generates high-contrast, movie-like scenes featuring the same consistent actor.

Across all these diverse backbones, RealDiffusion successfully disentangles the narrative generation from identity preservation, demonstrating that our physics-based prior is model-agnostic and highly versatile.

C.3. Impact of Identity Variation

To further verify the robustness of our disentangled architecture, we evaluated the model's ability to handle distinct identity inputs under identical narrative constraints. We selected two groups of contrasting identities and generated story sequences using fixed frame prompts for each group.

The results are presented in Figure 7. In the first two rows, despite the significant visual differences between the reference identities, RealDiffusion accurately reflects the unique facial features and appearance defined by the respective identity prompts. Crucially, the narrative progression—including specific poses, actions, and background elements—remains consistent across the different subjects. This confirms that our framework successfully disentangles character identity from narrative evolution. The insulated heat diffusion mechanism ensures that the specific identity features are propagated smoothly across frames without interfering with the structural layout dictated by the story prompts.

D. Long Story Generation

Generating coherent narratives over extended sequences poses a significant challenge for autoregressive-style or multi-frame diffusion models due to the problem of error accumulation. In many existing approaches, minor inconsistencies in early frames tend to propagate and amplify over time, causing the identity of the character to gradually drift away from the initial reference.

To evaluate the long-term stability of RealDiffusion, we generated a continuous story consisting of 40 frames. As visualized in Figure 8, we display the sequence in five consecutive segments covering the full progression from Frame 1 to Frame 40.

The results demonstrate that our framework maintains remarkable identity consistency throughout the vast majority of the 40-frame sequence. The insulated heat diffusion mechanism effectively acts as a continuous stabilizer by suppressing the high-frequency noise that typically leads to attribute drift. Consequently, the key features of the subject such as facial structure, hairstyle, and clothing details remain stable across the extended narrative. This stability persists even as the poses and scene contexts evolve dy-

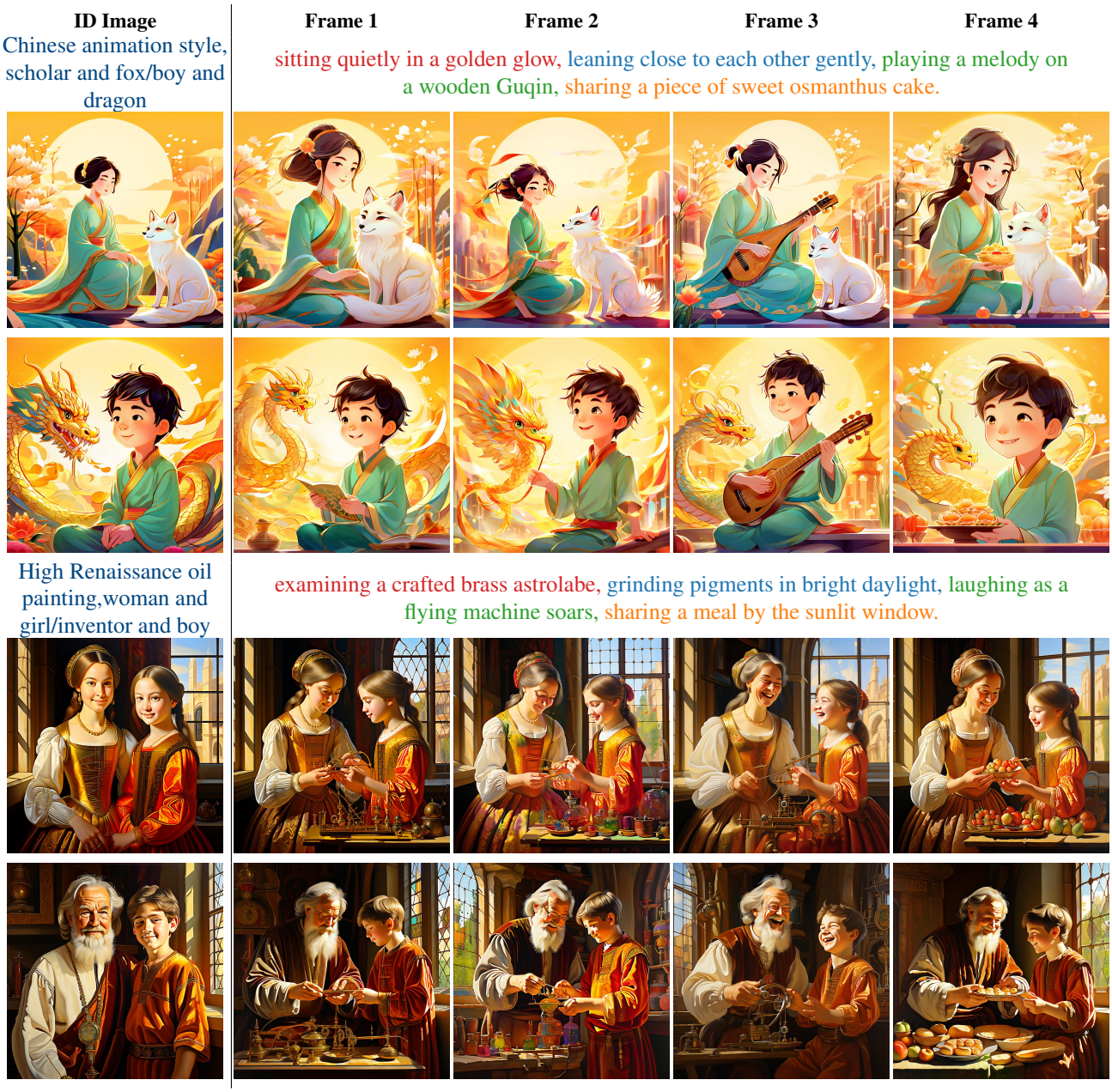

Figure 7. Qualitative evaluation of identity variation under fixed narrative prompts. Rows 1-2 and 3-4 share identical story descriptions but distinct identity inputs to demonstrate robust disentanglement.

namically according to the plot. This capability confirms that RealDiffusion is well-suited for long-form visual storytelling tasks where sustained coherence is paramount.

References

[1] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024. 5

[2] Mingxiao Li, Mang Ning, and Marie-Francine Moens. Consistent story generation: Unlocking the potential of zigzag sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[3] 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. 1

[4] Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fa-

Figure 8. A qualitative demonstration of long-term stability in a continuous 40-frame story generation task.

had 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

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

[6] 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

[7] 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

[8] 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. 1

[9] 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