Infinite-Story: A Training-Free Consistent Text-to-Image Generation Appendix

Evaluation Details

Evaluation protocol

To ensure consistency and comparability across models, we follow the evaluation protocol established by 1Prompt1Story (Liu et al. 2025), with an additional criterion of *style consistency*. The evaluation covers three key aspects: *prompt fidelity*, *identity consistency*, and *style consistency*.

Prompt Fidelity (CLIP-T). To assess *prompt fidelity*, we compute the CLIP text similarity (CLIP-T) using CLIP ViT-B/32 (Radford et al. 2021). Following (Liu et al. 2025), we prepend the prefix *"A photo depicts"* to each prompt and scale the cosine similarity by a factor of 2.5. The CLIP-T score is computed between the generated image and its paired prompt, and the final score is obtained by averaging across all samples.

Identity Consistency (CLIP-I and DreamSim). We measure *identity consistency* using two metrics: CLIP image similarity (CLIP-I) and DreamSim (Fu et al. 2023). Both are computed as the average pairwise similarity between images generated from the same identity prompt. To remove background bias, we apply CarveKit (Selin 2023) to extract the foreground subject and replace the background with uniform random noise, as done in (Liu et al. 2025). CLIP-I uses the ViT-B/16 model, and DreamSim provides perceptual similarity scores aligned with human judgment.

Style Consistency (DINO). To assess *style consistency* among images conditioned on the same identity prompt, we follow prior work on style-aligned image generation (Hertz et al. 2024; Park et al. 2025; Frenkel et al. 2024) and compute the average pairwise DINO similarity. Specifically, we use the CLS token from the DINO ViT-B/8 (Caron et al. 2021) model to measure global visual similarity, capturing alignment in rendering, background, and texture.

Harmonic Score(S_H). To jointly evaluate consistency across all dimensions, we report a harmonic score S_H defined as:

$$S_H = \text{HM}(\text{CLIP-T}, \text{CLIP-I}, 1 - \text{DreamSim}, DINO), \quad (5)$$

where HM denotes Harmonic Mean. This combined metric penalizes inconsistency in any single component, providing a robust measure of overall generation quality.

Implementation. We adopt the official evaluation scripts from (Liu et al. 2025), with minor adaptations (Add DINO metric). All metrics are computed on a single A6000 GPU using PyTorch. Background removal is applied consistently for identity-based metrics, and all features are extracted following standard preprocessing pipelines provided by each model.

Image-based consistent text-to-image models

We compare our method against several image-based consistent text-to-image generation approaches that leverage Stable Diffusion XL (Podell et al. 2023) as their backbone. Specifically, we include the following representative models:

- **IP-Adapter** (Ye et al. 2023): We use the official code.
- **PhotoMaker** (Li et al. 2024): We use the official code.
- **StoryDiffusion** (Zhou et al. 2024b): We use the official repository.
- **OneActor** (Wang et al. 2024): We use the official repository.

For all methods, we adopt the default DDIM sampling settings provided in their open-source implementations. During inference, each of these models requires an external reference image as an additional input (one or more). Therefore, we generate the reference image by providing only the identity portion of the full prompt to the corresponding base model. For instance, given the prompt "A graceful unicorn galloping through a flower field," we generate the reference image using "A graceful unicorn." This image is then used consistently across all prompts in the same sequence.

Non-reference consistent text-to-image models

We also compare our method against non-reference consistent text-to-image generation approaches that leverage Stable Diffusion XL (Podell et al. 2023) as their backbone. Specifically, we include the following representative models:

- **The chosen one** (Ye et al. 2023): We use the unofficial code.
- **ConsiStory** (Tewel et al. 2024): We use the official code.
- **1Prompt1Story** (Wang et al. 2024): We use the official repository.

For all methods, we adopt the default DDIM sampling settings provided in their open-source implementations. To ensure consistency across comparisons, we fix the number of DDIM sampling steps to 50 for all models, including the unofficial implementation of The Chosen One.

Details of User Study

To complement our quantitative evaluation, we conducted a user study with 50 participants, aged between 20 and 50. Each participant was shown a prompt along with four sets of generated images, each corresponding to a different method: our Infinite-Story, 1Prompt1Story (Liu et al. 2025), OneActor (Wang et al. 2024), and IP-Adapter (Ye et al. 2023).

Participants were instructed to select the image set that best satisfied each of the following criteria:

Method	$S_H \uparrow$	DINO ↑	CLIP-T ↑	CLIP-I ↑	DreamSim ↓
Vanilla Switti (Voronov et al. 2024)	0.7719	0.6595	**0.8904**	0.8871	0.2934
Switti + Ours	**0.8146**	**0.7441**	0.8756	**0.9018**	**0.2398**
Vanilla HART (Tang et al. 2024)	0.7434	0.6381	**0.8848**	0.8714	0.3488
HART + Ours	**0.7894**	**0.7048**	0.8505	**0.8982**	**0.2945**

Table 4: Effect of our technique on other scale-wise autoregressive model family.

- **Identity consistency**: Please select the option, from Option 1 to Option 4, that you find to have the most consistent appearance of the subject.
- **Prompt fidelity**: Please select the option, from Option 1 to Option 4, that you find to have the most consistent style throughout the image set.
- **Style consistency**: Please select the option, from Option 1 to Option 4, that you think best matches the text description.

Each participant evaluated multiple sets across diverse prompts in randomized order. An example of the interface used in the study is shown in Figure 9.

Comprehensive Analysis of Identity Prompt Replacement (IPR)

To further illustrate the effect of the Identity Prompt Replacement (IPR) module, we present a qualitative analysis of its influence on the generated images. Rather than copying the exact visual appearance of the object across different scenes, IPR helps preserve essential semantic attributes, such as gender and age, across all frames, as shown in Figure 8. This alignment of intrinsic characteristics plays a key role in enhancing the qualitative perception of coherence across the generated images. For instance, depending on the attributes of the reference instance, Figure 8-(top) shows that the baseline model tends to generate an older-looking object with a beard, whereas IPR adjusts the age to better match that of the reference. Similarly, Figure 8 demonstrates that IPR aligns the gender of the generated object with the reference, which is especially evident in the yellow boxes in the fourth column.

Generality of our method

To demonstrate the generalization capability, we applied our method to other scale-wise autoregressive text-to-image (T2I) generation models, Switti (Voronov et al. 2024) and HART (Tang et al. 2024). As shown in Table 4, our method demonstrates clear performance improvements across both models, especially in DINO, CLIP-I, and DreamSim. These results highlight the proposed technique's generalization capability beyond the infinity model and are applicable to a broader class of scale-wise autoregressive architectures.

Figure 8: Qualitative analysis of Identity Prompt Replacement (IPR).

Additional Ablation study on Adaptive Style Injection

We conduct an additional ablation study on the scaling coefficient λ used in Adaptive Style Injection. This experiment aims to analyze how different values of λ influence each evaluation metric. As shown in Table 5, increasing λ generally improves identity and style consistency metrics such as DINO, CLIP-I, and DreamSim. Notably, the best DINO and DreamSim scores are achieved when λ=0.9, indicating strong identity and style consistency. However, CLIP-T, which measures prompt fidelity, tends to degrade as λ increases, achieving its highest value at λ = 0.6. To strike a better balance between prompt fidelity and consistency, we select λ=0.85 as our default, which outperforms λ=0.9 in CLIP-T (0.8732 vs. 0.8722) while still maintaining competitive performance in other metrics.

Limitation of Infinite-story

Our Infinite-Story relies on a single reference image (anchor) within each batch to propagate identity and style features. While this enables efficient and training-free inference, it introduces sensitivity to anchor selection. If the anchor image is of low quality or stylistically off-target, this degradation may propagate to the entire batch. Notably, as our method does not alter the generation capabilities of the underlying Infinity model, its success is inherently tied to the

Parameter	$S_H \uparrow$	DINO ↑	CLIP-T ↑	CLIP-I ↑	DreamSim ↓
$\lambda = 0.6$	0.8420	0.7967	**0.8745**	0.9209	0.1998
$\lambda = 0.7$	0.8473	0.7865	<u>0.8737</u>	0.9227	0.1919
$\lambda = 0.8$	0.8506	0.8058	0.8735	0.9245	0.1904
$\lambda = 0.85$ (Ours)	**0.8538**	<u>0.8089</u>	0.8732	**0.9267**	<u>0.1834</u>
$\lambda = 0.9$	**0.8538**	**0.8102**	0.8722	<u>0.9251</u>	**0.1826**

Table 5: Ablation study on the Adaptive Style Injection scaling coefficient λ. Symbols ↑ and ↓ indicate whether higher or lower values are better. **Bold** and <u>underline</u> denote the best and second-best results, respectively.

quality of the initial output. This highlights the importance of future work in developing adaptive anchor selection or correction mechanisms.

Long story generation

To demonstrate the effectiveness of our method in generating extended, coherent visual narratives, we present two long-form examples in Figure 10 and Figure 11. Each figure illustrates a 10-frame story, where each image is conditioned on a unique prompt that reflects fine-grained scene variations while maintaining identity and style consistency.

Figure 10 showcases a fantasy narrative titled *"The leprechaun's Quest for His Lost Gold"*, where a leprechaun performs a series of playful actions leading to the discovery of a rainbow. The consistent character appearance and stylistic rendering across dynamic scenes demonstrate the model's ability to retain coherence in identity and global style throughout an extended sequence.

Figure 11 presents a slice-of-life story titled *"From Sleepy Eyes to a Warm Cup of Coffee"*, following a woman's transition from waking up to enjoying her coffee. Despite diverse poses, expressions, and indoor/outdoor transitions, the generated images preserve a unified visual identity and aesthetic style.

These results highlight Infinite-Story's ability to support rich, multi-prompt storytelling applications such as comic strips, storyboarding, and animated content creation—all while preserving consistency across identity, style, and prompt fidelity.

Addtional qualitative results

We present additional qualitative results of our Infinite-Story in Figure 12 and Figure 13. These additional results demonstrate that our method successfully preserves both identity and style consistency across diverse scenarios, while accurately reflecting the given text prompt. These results highlight the potential for broader practical applicability in various fields.

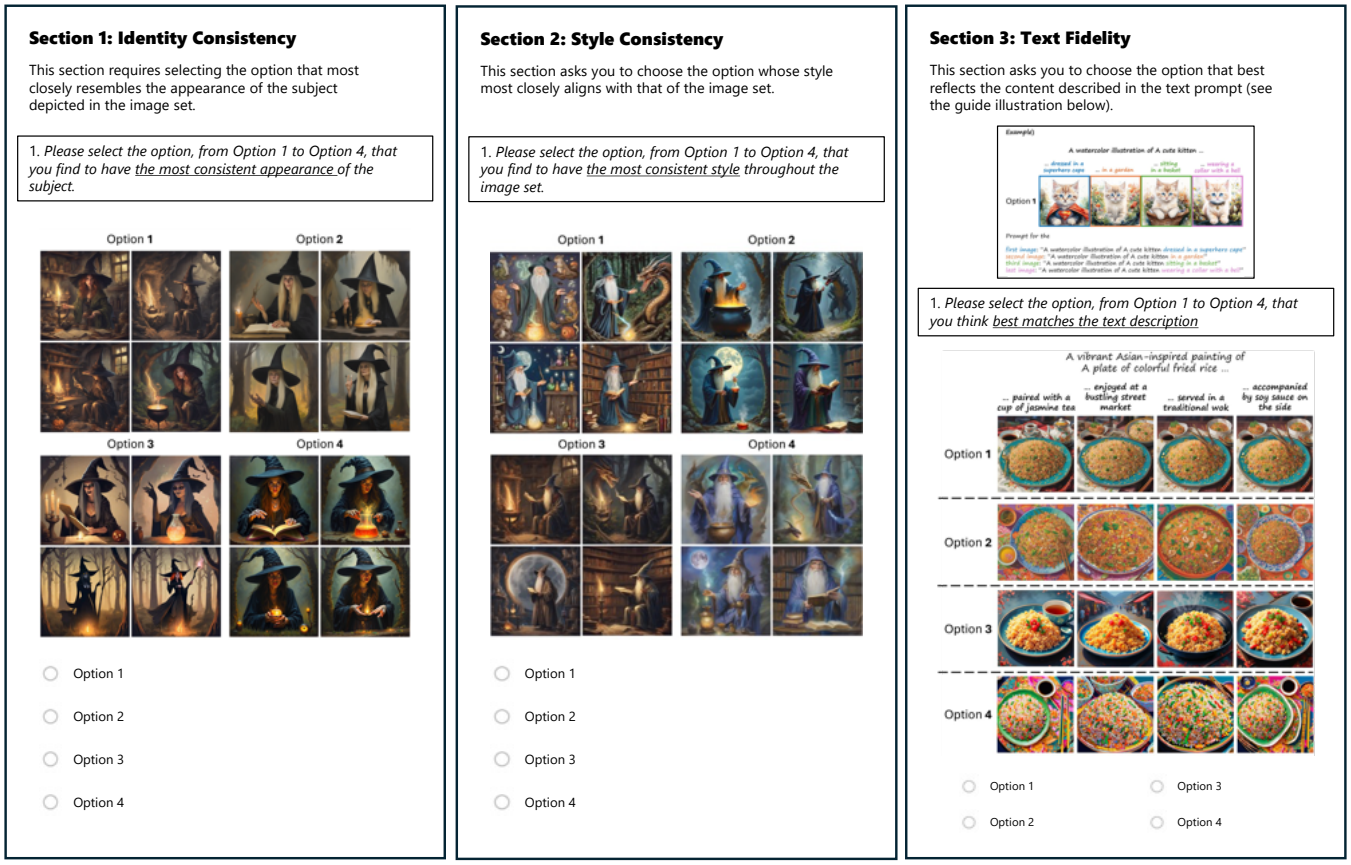

Figure 9: Example interface used in the user study. Participants selected the best-performing method among four candidates for each evaluation criterion.

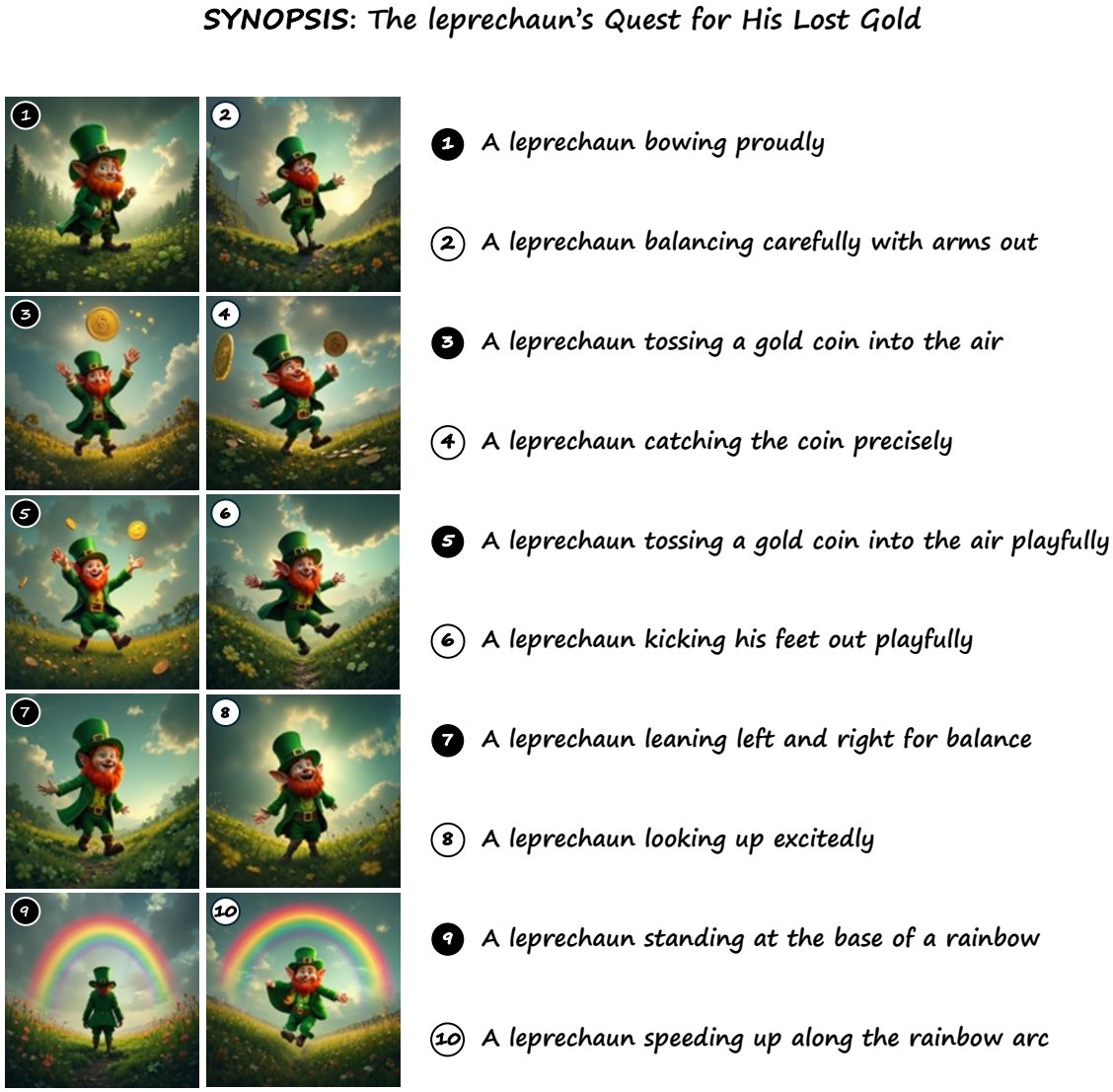

Figure 10: Long story generated by Infinite-Story. Under a unified synopsis, the leprechaun's quest for finding gold coin unfolds through organically connected scenes while preserving consistency of identity and style.

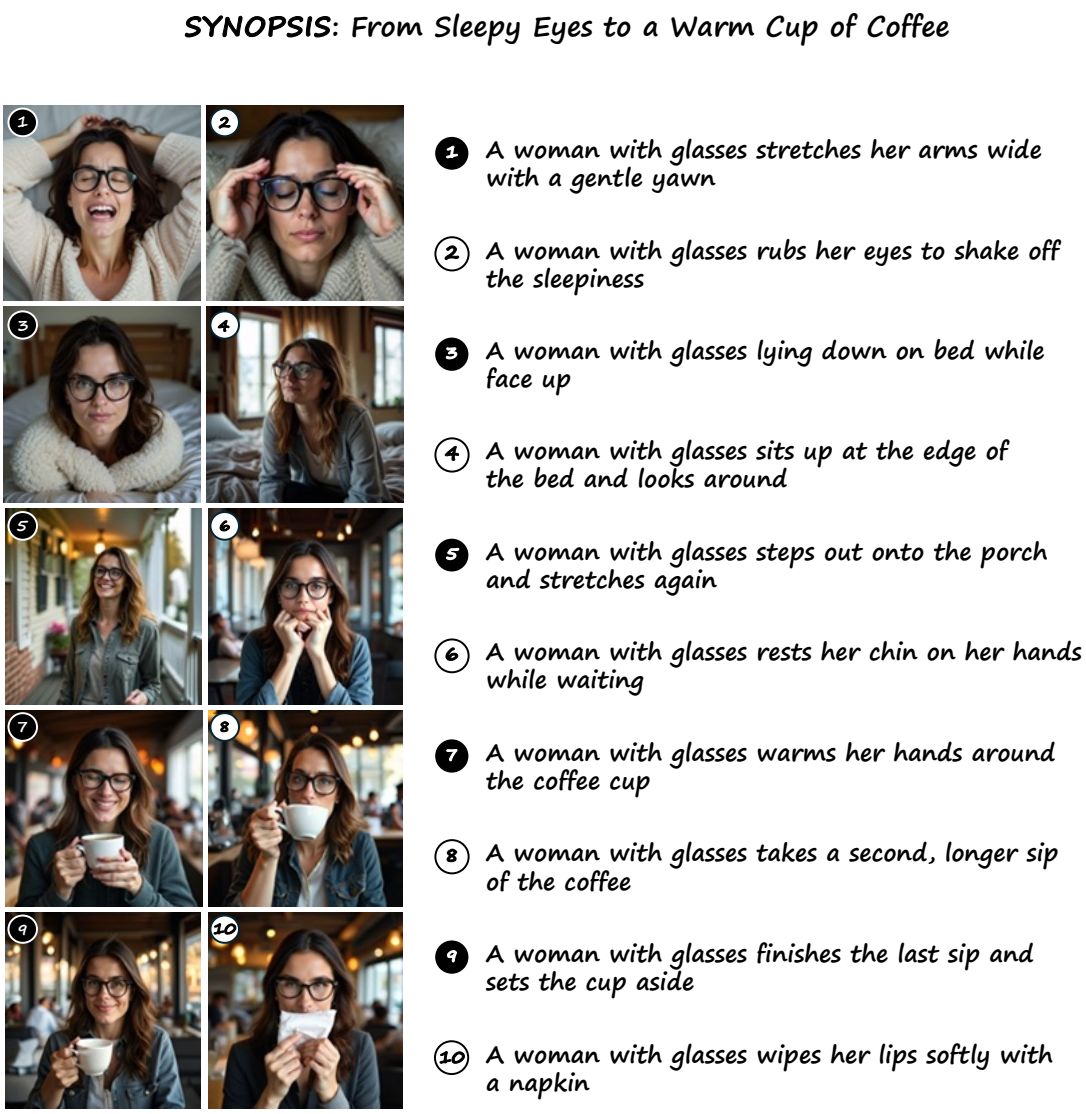

Figure 11: Long story generated by Infinite-Story. Under a unified synopsis, the story of woman with glasses transitions from waking up to enjoying her coffee at cafe unfolds through organically connected scenes while preserving consistency of identity and style.

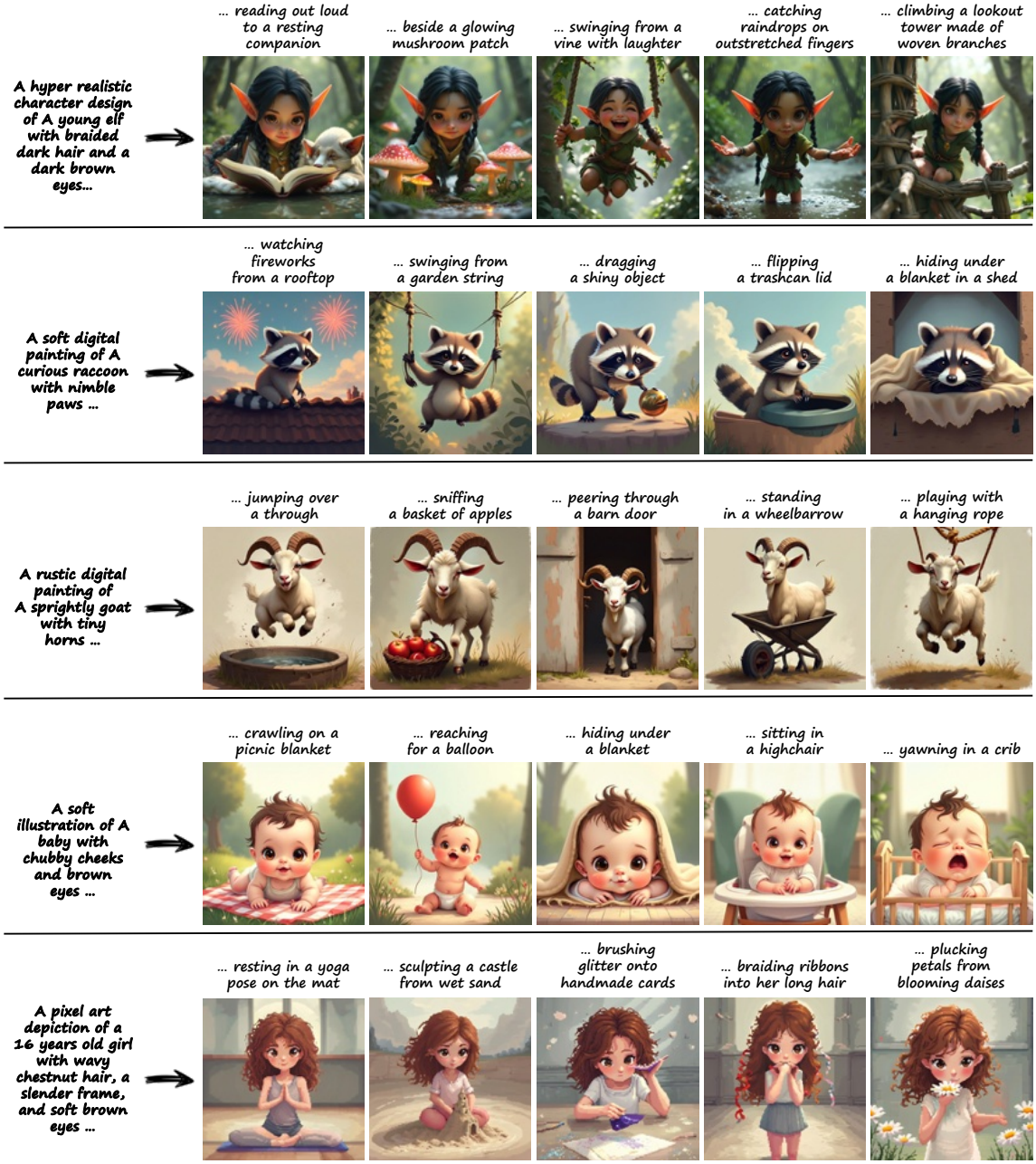

Figure 12: Additional qualitative results of Infinite-Story. Each row presents a set of images generated with a shared identity prompt combined with diverse expression prompts.

Figure 13: Additional qualitative results of Infinite-Story. Each row presents a set of images generated with a shared identity prompt combined with diverse expression prompts.