ONE-PROMPT-ONE-STORY: FREE-LUNCH CONSISTENT TEXT-TO-IMAGE GENERATION USING A SINGLE PROMPT

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generation models can create high-quality images from input prompts. However, they struggle to support the consistent generation of identity-preserving requirements for storytelling. Existing approaches to this problem typically require extensive training in large datasets or additional modifications to the original model architectures. This limits their applicability across different domains and diverse diffusion model configurations. In this paper, we first observe the inherent capability of language models, coined *context consistency*, to comprehend identity through context with a single prompt. Drawing inspiration from the inherent *context consistency*, we propose a novel *training-free* method for consistent text-to-image (T2I) generation, termed "One-Prompt-One-Story" (*1Prompt1Story*). Our approach *1Prompt1Story* concatenates all prompts into a single input for T2I diffusion models, initially preserving character identities. We then refine the generation process using two novel techniques: *Singular-Value Reweighting* and *Identity-Preserving Cross-Attention*, ensuring better alignment with the input description for each frame. In our experiments, we compare our method against various existing consistent T2I generation approaches to demonstrate its effectiveness through quantitative metrics and qualitative assessments.

1 INTRODUCTION

Text-based image generation (T2I) (Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022) aims to generate high-quality images from textual prompts, depicting various subjects in various scenes. The ability of T2I diffusion models to maintain *subject consistency* across a wide range of scenes is crucial for applications such as animation (Hu, 2024; Guo et al., 2024), storytelling (Yang et al., 2024; Gong et al., 2023; Cheng et al., 2024), video generation models (Khachatryan et al., 2023; Blattmann et al., 2023) and other narrative-driven visual applications. However, achieving consistent T2I generation remains a challenge for existing models, as shown in Fig. 1 (up).

Recent studies tackle the challenge of maintaining subject consistency through diverse approaches. Most methods require time-consuming training on large datasets for clustering identities (Avrahami et al., 2023), learning large mapping encoders (Gal et al., 2023b; Ruiz et al., 2024), or performing fine-tuning (Ryu, 2023; Kopiczko et al., 2024), which carries the risk of inducing language drift (Heng & Soh, 2024; Wu et al., 2024a; Huang et al., 2024), etc. Several recent training-free approaches (Tewel et al., 2024; Zhou et al., 2024) demonstrate remarkable results in generating images with consistent subjects by leveraging shared internal activations from the pre-trained models. These methods require extensive memory resources or complex module designs to strengthen the T2I diffusion model to generate satisfactory consistent images. However, they all neglect the inherent property of long prompts that identity information is implicitly maintained by context understanding, which we refer to as the *context consistency* of language models. For example, the dog object in "A dog is watching the movie. Afterward, the dog is lying in the garden." can be easily understood as the same without any confusion since it appears in the same paragraph and is connected by the context. We take advantage of this inherent feature to eliminate the requirement of additional finetuning or complicated module design.

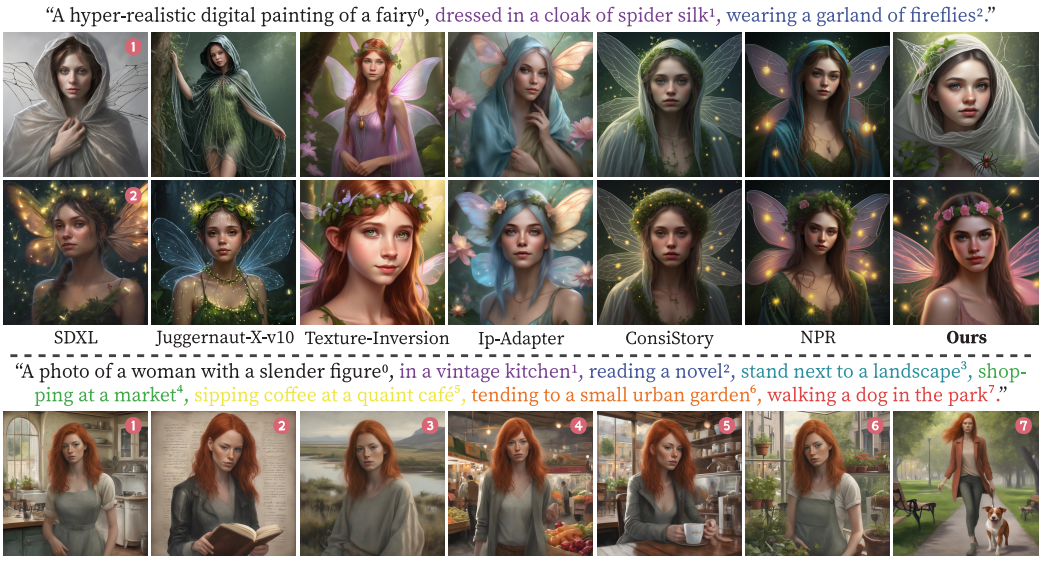

Figure 1: **Existing methods (up)** encounter challenges in consistent T2I generation. T2I models such as SDXL (Podell et al., 2023) and Juggernaut-X-v10 (RunDiffusion, 2024) often exhibit noticeable identity *inconsistency* across generated images. Although recent methods including IP-Adapter and ConsiStory have improved *identity consistency*, they lost the alignment between the generated images and corresponding input prompts. **Additional results of our *1Prompt1Story* (down)** demonstrate superior consistency without compromising the alignment between text and images.

Observing the inherent *context consistency* of language models, we propose a novel approach to generate images with consistent characters using a single prompt, termed *One-Prompt-One-Story* (*1Prompt1Story*). Specifically, *1Prompt1Story* consolidates all desired prompts into a single longer sentence, which starts with an *identity prompt* that describes the corresponding identity attributes and continues with subsequent *frame prompts* describing the desired scenarios in each frame. We denote this first step as *prompts consolidation*. By reweighting the consolidated prompt embeddings, we can easily implement a basic method *Naive Prompt Reweighting* to adjust the T2I generation performance, and this approach inherently achieves excellent identity consistency. Fig. 1 (up, the 6th column) illustrates two examples, each featuring an image generated with different frame descriptions within a single prompt by reweighting the frame prompt embeddings. These examples demonstrate that *Naive Prompt Reweighting* is able to maintain identity consistency with various scenario prompts. However, this basic approach does not guarantee strong text-image alignment for each frame, as the semantics of each frame prompt are usually intertwined within the consolidated prompt embedding (Radford et al., 2021). To further enhance text-image alignment and identity consistency of the T2I generative models, we introduce two additional techniques: *Singular-Value Reweighting* (SVR) and *Identity-Preserving Cross-Attention* (IPCA). The *Singular-Value Reweighting* aims to refine the expression of the prompt of the current frame while attenuating the information from the other frames. Meanwhile, the strategy *Identity-Preserving Cross-Attention* strengthens the consistency of the subject in the cross-attention layers. By applying our proposed techniques, *1Prompt1Story* achieves more consistent T2I generation results compared to existing approaches.

In the experiments, we extend an existing consistent T2I generation benchmark as *ConsiStory+* and compare it with several state-of-the-art methods, including ConsiStory (Tewel et al., 2024), StoryDiffusion (Zhou et al., 2024), IP-Adapter (Ye et al., 2023), etc. Both qualitative and quantitative performance demonstrate the effectiveness of our method *1Prompt1Story*. In summary, the main contributions of this paper are:

- To the best of our knowledge, we are the first to analyze the overlooked ability of language models to maintain inherent *context consistency*, where multiple frame descriptions within a single prompt inherently refer to the same subject identity.
- Based on the *context consistency* property, we propose *One-Prompt-One-Story* as a novel *training-free* method for consistent T2I generation. More specifically, we further propose *Singular-Value Reweighting* and *Identity-Preserving Cross-Attention* techniques to improve text-image alignment and subject consistency, allowing each frame prompt to be individually expressed within a single prompt while maintaining a consistent identity along with the *identity prompt*.

- Through extensive comparisons with existing consistent T2I generation approaches, we confirm the effectiveness of *1Prompt1Story* in generating images that consistently maintain identity throughout a lengthy narrative over our extended *ConsiStory+* benchmark.

2 Related Work

T2I personalized generation. T2I personalization is also referred to *T2I model adaptation*. This aims to adapt a given model to a *new concept* by providing a few images and binding the new concept to a unique token. As a result, the adaptation model can generate various renditions of the new concept. One of the most representative methods is DreamBooth (Ruiz et al., 2023), where the pre-trained T2I model learns to bind a modified unique identifier to a specific subject given a few images, while it also updates the T2I model parameters. Recent approaches (Kumari et al., 2023; Han et al., 2023b; Shi et al., 2023) follow this pipeline and further improve the quality of the generation. Another representative, Textual Inversion (Gal et al., 2023a), focuses on learning new concept tokens instead of fine-tuning the T2I generative models. Textual Inversion finds new pseudo-words by conducting personalization in the text embedding space. The coming works (Dong et al., 2022; Voynov et al., 2023; Han et al., 2023a; Zeng et al., 2024) follow similar techniques.

Consistent T2I generation. Despite recent advances, T2I personalization methods often require extensive training to effectively learn modifier tokens. This training process can be time-consuming, which limits their practical impact. More recently, there has been a shift towards developing consistent T2I generation approaches (Wang et al., 2024b;a), which can be considered a specialized form of T2I personalization. These methods mainly focus on generating human faces that possess semantically similar attributes to the input images. Importantly, they aim to achieve this identity-preserving T2I generation without the need for additional fine-tuning. They mainly take advantage of PEFT techniques (Ryu, 2023; Kopiczko et al., 2024) or pre-training with large datasets (Ruiz et al., 2024; Xiao et al., 2023) to learn the image encoder to be customized in the semantic space. For example, PhotoMaker (Li et al., 2023b) enhances its ability to extract identity embeddings by fine-tuning part of the transformer layers in the image encoder and merging the class and image embeddings. The Chosen One (Avrahami et al., 2023) utilizes an identity clustering method to iteratively identify images with a similar appearance from a set of images generated by identical prompts.

However, most consistent T2I generation methods (Akdemir & Yanardag, 2024; Wang et al., 2024a) still require training the parameters of the T2I models, sacrificing compatibility with existing pre-trained community models, or fail to ensure high face fidelity. Additionally, as most of these systems (Li et al., 2023b; Gal et al., 2023b; Ruiz et al., 2024) are designed specifically for human faces, they encounter limitations when applied to non-human subjects. Even for the state-of-the-art approaches, including StoryDiffusion (Zhou et al., 2024), The Chosen One (Avrahami et al., 2023) and ConsiStory (Tewel et al., 2024), they either require time-consuming iterative clustering or high memory demand in generation to achieve identity consistency.

Storytelling. Story generation (Li et al., 2019; Maharana et al., 2021), also referred to as storytelling, is one of the active research directions that is highly related to character consistency. Recent researches (Tao et al., 2024; Wang et al., 2023) have integrated the prominent pre-trained T2I diffusion models (Rombach et al., 2022; Ramesh et al., 2022) and the majority of these approaches require intense training over storytelling datasets. For example, Make-a-Story (Rahman et al., 2023) introduces a visual memory module designed to capture and leverage contextual information throughout the storytelling process. StoryDALL-E (Maharana et al., 2022) extends the story generation paradigm to story continuation, using DALL-E capabilities to achieve substantial improvements over previous GAN-based methodologies. Note that the story continuation shares similarities with consistent Text-to-Image generation by using reference images. However, current consistent T2I generation methods prioritize preserving human face identities, whereas story continuation involves supporting various subjects or even multiple subjects within the generated images.

In this paper, our proposed consistent T2I framework, *1Prompt1Story*, diverges significantly from previous approaches in storytelling and consistent T2I generation methods. We explore the inherent *context consistency* property in language models instead of finetuning large models or designing complex modules. Importantly, it is compatible with various T2I generative models, since the properties of the text model are independent of the specific generation model used as the backbone.

3 METHOD

Consistent T2I generation aims to generate a set of images depicting consistent subjects in different scenarios using a set of prompts. These prompts start with an *identity prompt*, followed by the *frame prompts* for each subsequent visualization frame. In this section, we first empirically show that different frame descriptions included in a concatenated prompt can maintain identity consistency due to the inherent *context consistency* property of language models. We examine this observation through comprehensive analyses in Sec. 3.1 and propose the basic *Naive Prompt Reweighting* pipeline of our method *1Prompt1Story*. Following that, to ensure that each frame description within the prompt is expressed individually while diminishing the impact of other *frame prompts*, we introduce *Singular-Value Reweighting* and *Identity-Preserving Cross-Attention* in Sec. 3.2. The illustration of *1Prompt1Story* is shown in Fig. 4 and Algorithm 1 in the Appendix.

3.1 CONTEXT CONSISTENCY

Latent Diffusion Models. We built our approach on the SDXL (Podell et al., 2023) model. This latent diffusion model consists of two primary components: an autoencoder (i.e., an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$) and a diffusion model (i.e., ϵ_θ with parameter θ). The model ϵ_θ is trained by the loss:

$$L_{LDM} := \mathbb{E}_{z_0 \sim \mathcal{E}(x), \epsilon \sim \mathcal{N}(0,1), t \sim \text{Uniform}(1,T)} \left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\xi(\mathcal{P}))\|_2^2 \right], \tag{1}$$

where ϵ_θ is a UNet, conditioning a latent input z_t, a timestep $t \sim \text{Uniform}(1, T)$, and a text embedding $\tau_\xi(\mathcal{P})$. More specifically, text-guided diffusion models generate an image from the textual condition as $\mathcal{C} = \tau_\xi(\mathcal{P}) \in \mathbb{R}^{M \times D}$, where the text embedding is with M tokens and each token has a feature dimension of D and τ_ξ is the CLIP text encoder (Radford et al., 2021)[1]. The cross-attention map is derived from $\epsilon_\theta(z_t, t, \mathcal{C})$. Let f_{z_t} represent the output of the feature map from the network ϵ_θ. We get a query matrix $Q = l_Q(f_{z_t})$ using the projection network l_Q. Similarly, given a textual embedding $\mathcal{C}$, we compute a key matrix $\mathcal{K} = l_\mathcal{K}(\mathcal{C})$ with projection network $l_\mathcal{K}$. Then the attention map is computed according to: $\mathcal{A}_t = softmax(Q \cdot \mathcal{K}^T / \sqrt{d})$ where d is the dimension of queries and key, and the cell $[\mathcal{A}_t]_{ij}$ defines the weight of the j-th token on the i-th token.

Problem Setups. In the T2I diffusion models, the text embedding $\mathcal{C} = \tau_\xi(\mathcal{P}) \in \mathbb{R}^{M \times D}$ is with M tokens. The M tokens contain a start token [SOT] , followed by $|\mathcal{P}|$ tokens corresponding to the prompt, and $M - |\mathcal{P}| - 1$ padding end tokens [EOT] . Previous consistent T2I generation works (Avrahami et al., 2023; Tewel et al., 2024; Zhou et al., 2024) generate images from a set of N prompts. This set of prompts starts with an *identity prompt* $\mathcal{P}_0$ that describes the relevant attribute of the subject and continues with multiple frame prompt $\mathcal{P}_i$, where $i = 1, \ldots, N$ describes each frame scenario. However, this separate generation pipeline ignores the inherent language property, i.e., the *context consistency*, by which identity is consistently ensured by the context information inherent in language models. This property stems from the self-attention mechanism within Transformer-based text encoders (Radford et al., 2021; Vaswani et al., 2017), which allows learning the interaction between phrases in the text embedding space.

In the following, we analyze the *context consistency* under different prompt configurations in both textual space and image space. Specifically, we refer to the conventional prompt setups as *multi-prompt generation*, which is commonly used in existing consistent T2I generation methods. The multi-prompt generation uses N prompts separately for each generated frame, each sharing the same *identity prompt* and the corresponding frame prompt as $[\mathcal{P}_0; \mathcal{P}_i], i \in [1, N]$. In contrast, our *single-prompt generation* concatenates all the prompts as $[\mathcal{P}_0; \mathcal{P}_1; \ldots; \mathcal{P}_N]$ for each frame generation, which we refer as the *Prompt Consolidation (PCon)*.

3.1.1 CONTEXT CONSISTENCY IN TEXT EMBEDDINGS

Empirically, we find that the frame prompt $\{\mathcal{P}_i \mid i = 1, \ldots, N\}$ in the *single-prompt generation* setup have relatively small semantic distances among each other in the textual embedding space, whereas those across *multi-prompt generation* have comparatively larger distances. For instance, we set the identity frame $\mathcal{P}_0$ = "A watercolor of a cute kitten" as an example. We then create

[1] SDXL uses two text encoders and concatenate the embeddings as the final input. $M = 77$ by default.

$N = 5$ *frame prompts* $\{\mathcal{P}_i, i \in [1, N]\}$ as "in a garden", "dressed in a superhero cape", "wearing a collar with a bell", "sitting in a basket", and "dressed in a cute sweater", respectively. Under the multi-prompt setup, each frame is generated by the text embedding defined as $\mathcal{C}_i = \tau_\xi([\mathcal{P}_0; \mathcal{P}_i]) = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \boldsymbol{c}^{\mathcal{P}_i}, \boldsymbol{c}^{EOT}]$, $(i = 1, \ldots, N)$, while the text embedding of the *Prompt Consolidation* in the single-prompt case is $\mathcal{C} = \tau_\xi([\mathcal{P}_0; \mathcal{P}_1; \ldots; \mathcal{P}_N]) = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \boldsymbol{c}^{\mathcal{P}_1}, \ldots, \boldsymbol{c}^{\mathcal{P}_N}, \boldsymbol{c}^{EOT}]$.

To analyze the distances among the *frame prompts*, we extract $\boldsymbol{c}^{\mathcal{P}_i}$ from $\mathcal{C}_i$ for multi-prompt setup and apply t-SNE for 2D visualization (Fig. 2-left). Similarly, we extract all $\boldsymbol{c}^{\mathcal{P}_i}$ from $\mathcal{C}$ for the single-prompt setup (Fig. 2-left). As can be observed, the text embeddings of *frame prompts* under the multi-prompt setup are widely distributed in the text representation space (red dots) with an average Euclidean L_2 distance of 71.25. In contrast, the embeddings in the single-prompt case exhibit more compact distributions (blue dots), with a much smaller average L_2 distance of 46.42. We also performed a similar distance analysis on all prompt sets in our benchmark *ConsiStory+*. As shown in Fig.2-right, we can conclude a similar observation that the *frame prompts* share more similar semantic information and identity consistency within the single-prompt setup.

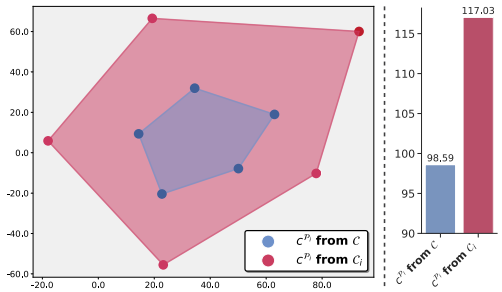

Figure 2: **t-SNE visualization of text embeddings (Left):** $\boldsymbol{c}^{\mathcal{P}_i}$ from *single-prompt generation* are closer together compared to those from *multi-prompt generation*. **Statistical results (Right):** We evaluated the average distances between the corresponding point sets of all prompt sets on the *ConsiStory+* benchmark after dimensionality reduction. The average distance between text embeddings from *single-prompt generation* is smaller than that from *multi-prompt generation*.

3.1.2 Context Consistency in Image Generation

To demonstrate that *context consistency* is also maintained in the image space, we further conducted image generation experiments using the prompt example above. The images generated by the SDXL model with the multi-prompt configuration, as illustrated in Fig. 3 (left, the first row), show various characters that lack identity consistency. Instead, we use our proposed concatenated prompt $\mathcal{P} = [\mathcal{P}_0; \mathcal{P}_1; \ldots; \mathcal{P}_N]$. To generate the i-th frame $(i = 1, ..., N)$, we reweight the $\boldsymbol{c}^{\mathcal{P}_i}$ corresponding to the desired frame prompt $\mathcal{P}_i$ by a magnification factor while rescaling the embeddings of the other *frame prompts* by a reduction factor. This modified text embedding is then imported to the T2I model to generate the frame image. We refer to this simplistic reweighting approach as *Naive Prompt Reweighting* (NPR). By this means, the T2I model synthesizes frame images with the same subject identity. However, the backgrounds get blended among these frames, as shown in Fig. 3 (left, the second row). By contrast, our full model *1Prompt1Story* introduced in Sec. 3.2 generates images with better consistent identity and text-image alignment for each frame prompt, as shown in Fig. 3 (left, the last row).

To visualize identity similarity among images, we removed backgrounds using CarveKit (Selin, 2023) and extracted visual features with DINO-v2 (Oquab et al., 2023; Darcet et al., 2023). These features are then projected into the 2D space by t-SNE (Hinton & Roweis, 2002) (as shown in Fig. 3 (mid)). Our complete approach *1Prompt1Story* obviously obtains better identity consistency than the other two comparison methods, while *Naive Prompt Reweighting* shows improvements over the SDXL baseline. We also applied the analysis across our extended benchmark *ConsiStory+* and calculated the average pairwise distance, as shown in Fig. 3 (right). These results further consolidate our conclusion that the *frame prompts* in a single-prompt setup share more identity consistency than the multi-prompt case.

3.2 One-Prompt-One-Story

As also observed from the above section, simply concatenating the prompts as *Naive Prompt Reweighting* cannot guarantee that the generated images accurately reflect the frame prompt descriptions, for which we assume that the T2I model cannot accurately capture the correct partition of the concatenated prompt embeddings. Furthermore, the various semantics within the consolidated descriptions interact with each other (Chefer et al., 2023; Rassin et al., 2024). To mitigate

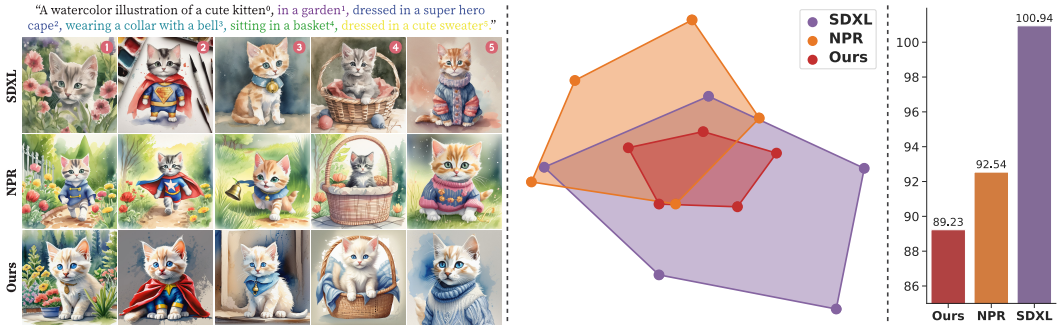

Figure 3: **(Left):** SDXL generates frame images using multi-prompt generation, while *Naive Prompt Reweighting* (NPR) and our method utilize the single-prompt setup. **(Mid):** Image features are extracted by DINO-v2 (Oquab et al., 2023) and visualized by the t-SNE reduction. *Naive Prompt Reweighting* and *1Prompt1Story* show more consistent identity generations than the SDXL model. **(Right):** Statistics of the average feature distances among generated images from the prompts in our extended *ConsiStory+* benchmark, which further confirms that *1Prompt1Story* produces better identity consistency.

this issue, we propose additional techniques based on the *Prompt Consolidation (PCon)*, namely *Singular-Value Reweighting* (SVR) and *Identity-Preserving Cross-Attention* (IPCA).

Singular-Value Reweighting. After the *Prompt Consolidation* as $\mathcal{C} = \tau_\xi([\mathcal{P}_0; \mathcal{P}_1; \ldots; \mathcal{P}_N]) = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \boldsymbol{c}^{\mathcal{P}_1}, \ldots, \boldsymbol{c}^{\mathcal{P}_N}, \boldsymbol{c}^{EOT}]$, we require the current frame prompt to be better *expressed* in the T2I generation, which we denote as $\mathcal{P}^{exp} = \mathcal{P}_j, (j = 1, ..., N)$. We also expect the remaining frames to be *suppressed* in the generation, which we denote as $\mathcal{P}^{sup} = \mathcal{P}_k, k \in [1, N] \backslash \{j\}$. Thus, the N *frame prompts* of the subject description can be written as $\{\mathcal{P}^{exp}, \mathcal{P}^{sup}\}$. As the [EOT] token contains significant semantic information (Li et al., 2023a; Wu et al., 2024b), the semantic information corresponding to $\mathcal{P}^{exp}$, in both $\mathcal{P}_j$ and [EOT] , needs to be enhanced, while the semantic information corresponding to $\mathcal{P}^{sup}$, in $\mathcal{P}_k$, $k \neq j$ and [EOT] , need to be suppressed. We extract the token embeddings for both express and suppress sets as $\mathcal{X}^{exp} = [\boldsymbol{c}^{\mathcal{P}_j}, \boldsymbol{c}^{EOT}]$ and $\mathcal{X}^{sup} = [\boldsymbol{c}^{\mathcal{P}_1}, \ldots, \boldsymbol{c}^{\mathcal{P}_{j-1}}, \boldsymbol{c}^{\mathcal{P}_{j+1}}, \ldots, \boldsymbol{c}^{\mathcal{P}_N}, \boldsymbol{c}^{EOT}]$.

Inspired by (Gu et al., 2014; Li et al., 2023a), we assume that the main singular values of $\mathcal{X}^{exp}$ correspond to the fundamental information of $\mathcal{P}^{exp}$. We then perform SVD decomposition as: $\mathcal{X}^{exp} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where $\boldsymbol{\Sigma} = diag(\sigma_0, \sigma_1, \cdots, \sigma_{n_j})$, the singular values $\boldsymbol{\sigma}_0 \geq \cdots \geq \boldsymbol{\sigma}_{n_j}$ [2]. To enhance the expression of the frame $\mathcal{P}_j$, we introduce the augmentation for each singular value, termed as **SVR+** and formulated as:

$$\hat{\sigma} = \beta e^{\alpha\sigma} * \sigma. \tag{2}$$

where the symbol e is the exponential, α and β are parameters with positive numbers. We recover the tokens as $\hat{\mathcal{X}}^{exp} = \boldsymbol{U}\hat{\boldsymbol{\Sigma}}\boldsymbol{V}^T$, with the updated $\hat{\boldsymbol{\Sigma}} = diag(\hat{\sigma_0}, \hat{\sigma_1}, \cdots, \hat{\sigma_{n_j}})$. The new prompt embedding is defined as $\hat{\mathcal{X}}^{exp} = [\hat{\boldsymbol{c}}^{\mathcal{P}_j}, \hat{\boldsymbol{c}}^{EOT}]$, and $\hat{\mathcal{C}} = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \cdots, \hat{\boldsymbol{c}}^{\mathcal{P}_j}, \cdots, \boldsymbol{c}^{\mathcal{P}_N}, \hat{\boldsymbol{c}}^{EOT}]$. Note that there is an updated $\hat{\mathcal{X}}^{sup} = [\boldsymbol{c}^{\mathcal{P}_1}, \ldots, \boldsymbol{c}^{\mathcal{P}_{j-1}}, \boldsymbol{c}^{\mathcal{P}_{j+1}}, \ldots, \boldsymbol{c}^{\mathcal{P}_N}, \hat{\boldsymbol{c}}^{EOT}]$.

Similarly, we suppress the expression of the remaining frames. Since $\hat{\mathcal{X}}^{sup}$ contains information related to multiple frames, the main singular values of SVD in $\hat{\mathcal{X}}^{sup}$ only capture a small portion of these descriptions, which may lead to insufficient weakening of such semantics (as shown in the Appendix of Fig. 11-right). Therefore, we propose to weaken each frame prompt in $\hat{\mathcal{X}}^{sup}$ separately. We construct the matrix as $\hat{\mathcal{X}}_k^{sup} = [\boldsymbol{c}^{\mathcal{P}_k}, \hat{\boldsymbol{c}}^{EOT}], k \neq j$ to perform SVD with the singular values $\hat{\boldsymbol{\sigma}}_0 \geq \cdots \geq \hat{\boldsymbol{\sigma}}_{n_k}$. Then, each singular value is weakened as follows, termed as **SVR-**:

$$\tilde{\sigma} = \beta' e^{-\alpha' \hat{\sigma}} * \hat{\sigma}. \tag{3}$$

where α' and β' are parameters with positive numbers. The recovered structure is $\tilde{\mathcal{X}}_k^{sup} = [\tilde{\boldsymbol{c}}^{\mathcal{P}_k}, \tilde{\boldsymbol{c}}^{EOT}]$, After reducing the expression of each suppress token, we finally obtain the new text embedding $\tilde{\mathcal{C}} = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \tilde{\boldsymbol{c}}^{\mathcal{P}_1}, \cdots, \hat{\boldsymbol{c}}^{\mathcal{P}_j}, \cdots, \tilde{\boldsymbol{c}}^{\mathcal{P}_N}, \tilde{\boldsymbol{c}}^{EOT}]$.

[2] $n_j = \min(D, |\boldsymbol{c}^{\mathcal{P}_j}| + |\boldsymbol{c}^{EOT}|)$. The dimension D in the SDXL model is greater than $|\boldsymbol{c}^{\mathcal{P}_j}| + |\boldsymbol{c}^{EOT}|$)

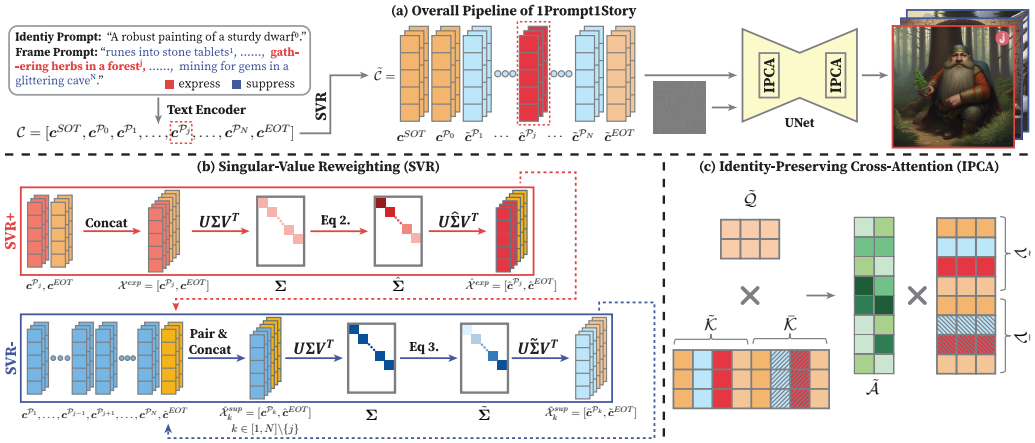

Figure 4: **(a):** The overall pipeline of *1Prompt1Story*. We combine the *identity prompt* and *frame prompts* into a single prompt, then we apply both *Singular-Value Reweighting* (SVR) and *Identity-Preserving Cross-Attention* (IPCA) to generate identity-consistent images. **(b):** During *SVR*, we first enhance the semantic information of the *express set* $\mathcal{X}^{exp}$ (red arrow), then iteratively weaken the semantics for the *suppress set* $\mathcal{X}^{sup}$ (blue arrow). **(c):** In *IPCA*, we concatenate $\tilde{\mathcal{K}}$ with $\bar{\mathcal{K}}$ and $\tilde{\mathcal{V}}$ with $\bar{\mathcal{V}}$ to improve identity consistency.

Identity-Preserving Cross-Attention. The use of *Singular-Value Reweighting* can reduce the blending of frame descriptions in *single-prompt generation*. However, we observed that it can also impact *context consistency* within the single prompt, leading to images generated slightly less similar in identity (as shown in the ablation study of Fig. 7). Recent work (Liu et al., 2024) demonstrated that cross-attention maps capture the characteristic information of the token, while self-attention preserves the layout information and the shape details of the image. Inspired by this, we propose *Identity-Preserving Cross-Attention* to further enhance the identity similarity between images generated from the concatenated prompt of our proposed *Prompt Consolidation*.

For a specific timestep t, after applying *Singular-Value Reweighting*, we have the updated text embedding $\tilde{\mathcal{C}}$. During a denoising pass through the diffusion model, we obtain the corresponding $\tilde{\mathcal{Q}}, \tilde{\mathcal{K}}, \tilde{\mathcal{V}}$ in the cross-attention layer. Here, we aim to strengthen the identity consistency among the images and mitigate the impact of irrelevant prompts. We set the token features in $\tilde{\mathcal{K}}$ corresponding to $\mathcal{P}_i, i \in [1, N]$ to zero , resulting in $\bar{\mathcal{K}}$. Here, only the *identity prompt* remains to augment the identity semantics. Similarly, we can get $\bar{\mathcal{V}}$. We form a new version of $\tilde{\mathcal{K}}$ by concatenating it with $\bar{\mathcal{K}}$, dubbed $\tilde{\mathcal{K}} = \texttt{Concat}(\tilde{\mathcal{K}}^\top, \bar{\mathcal{K}}^\top)^\top$. The new cross-attention map is then given by:

$$\tilde{\mathcal{A}} = softmax\left(\tilde{\mathcal{Q}}\tilde{\mathcal{K}}^\top / \sqrt{d}\right) \tag{4}$$

where d is the dimension of $\tilde{\mathcal{Q}}$ and $\tilde{\mathcal{K}}$. Similarly, we update $\tilde{\mathcal{V}} = \texttt{Concat}(\tilde{\mathcal{V}}^\top, \bar{\mathcal{V}}^\top)^\top$. The final output feature of the cross-attention layer is $\tilde{\mathcal{A}} \times \tilde{\mathcal{V}}$. This output is a reweighted version that strengthens identity consistency using filtered features, which only contain the *identity prompt* semantics.

4 Experiments

4.1 Experimental Setups

Comparison Methods and Benchmark. We compare our method with the following consistent T2I generation approaches: BLIP-Diffusion (Li et al., 2024), Textual Inversion (TI)(Gal et al., 2023a), IP-Adapter(Ye et al., 2023), PhotoMaker (Li et al., 2023b), The Chosen One (Avrahami et al., 2023), ConsiStory (Tewel et al., 2024), and StoryDiffusion (Zhou et al., 2024). We follow the default configurations in their papers or open-source implementations.

To evaluate their performance, we introduce *ConsiStory+*, an extension of the original ConsiStory (Tewel et al., 2024) benchmark. This new benchmark incorporates a wider range of subjects, descriptions, and styles. Following the evaluation protocol outlined in ConsiStory, we evaluated both *prompt alignment* and *subject consistency* across *ConsiStory+*, generating up to 1500 images on 200

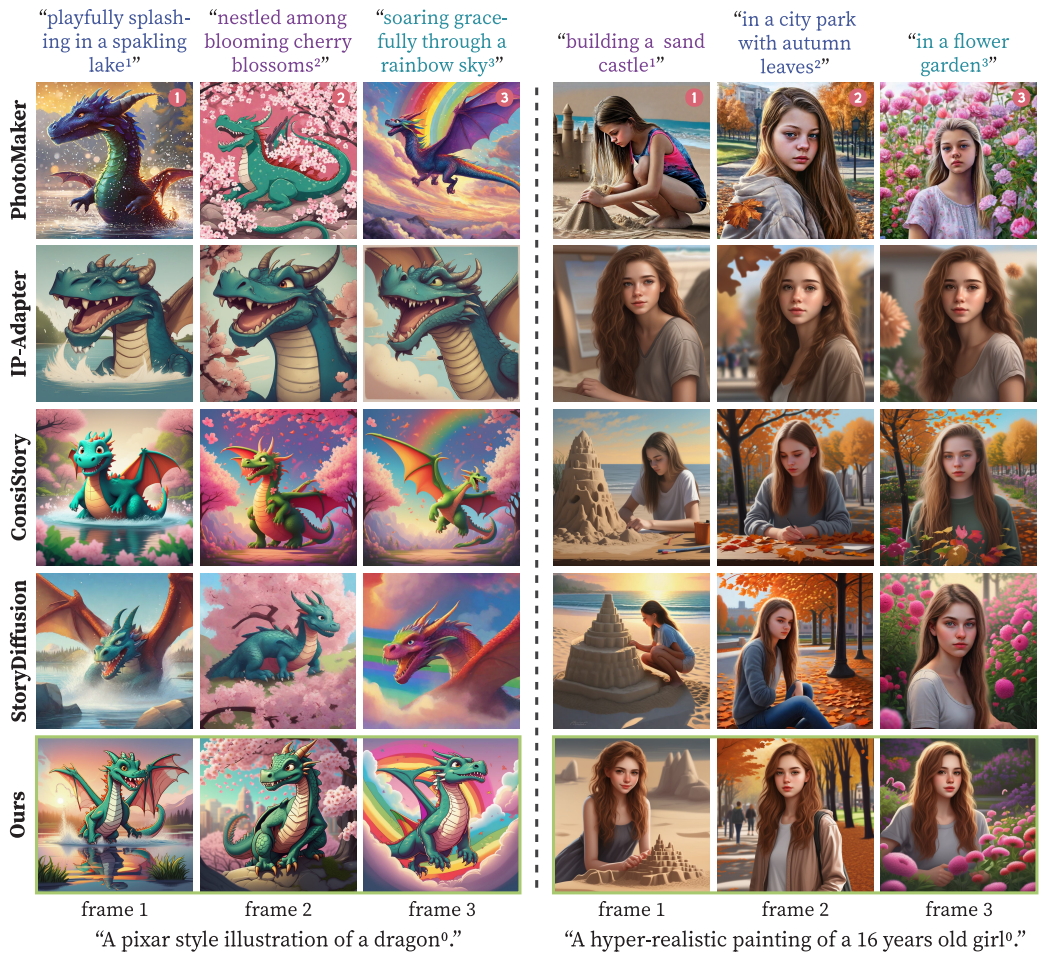

Figure 5: **Qualitative results**. We compare our method with PhotoMaker, IP-Adapter, ConsiStory, and StoryDiffsion. Among them, Texture Inversion, PhotoMaker, ConsiStory, and StoryDiffsion struggled to maintain identity consistency for the *dragon* object while IP-Adapter produced images with relatively similar poses and backgrounds. See Comparison with the remaining methods in Fig. 22 of the Appendix.

prompt sets. Additional details on the construction of our benchmark and the implementation of the methods are provided in Appendix B.2 and Appendix B.3.

Evaluation Metrics. To assess *prompt alignment* performance, we compute the average CLIP-Score (Hessel et al., 2021) for each generated image in relation to its corresponding prompt, which we denote as CLIP-T. For the *identity consistency* evaluation, we measure image similarity using both DreamSim (Fu et al., 2023), which has been shown to closely reflect human judgment in evaluating visual similarity, and CLIP-I (Hessel et al., 2021), calculated by the cosine distance between image embeddings. In line with the methodology proposed in DreamSim (Fu et al., 2023), we remove image backgrounds using CarveKit (Selin, 2023) and replace them with random noise to ensure that similarity measurements focus solely on the identities of subjects.

4.2 Experimental Results

Qualitative Comparison. In Fig. 5, we present the qualitative comparison results. Our method *1Prompt1Story* demonstrates well-balanced performance in several key aspects, including identity preservation, accurate frame descriptions, and diversity in the pose of objects. In contrast, other methods exhibit shortcomings in one or more of these aspects. Specifically, PhotoMaker, ConsiStory, and StoryDiffusion all produce inconsistent identities for the subject "dragon" in the examples on the left. Additionally, IP-Adapter tends to generate images with repetitive poses and similar backgrounds, often neglecting frame prompt descriptions. ConsiStory also displays duplicated background generation in the consistent T2I generation.

Table 1: **Quantitative comparison.** The best and second best results are highlighted in **bold** and underlined, respectively. Vanilla SD1.5 and Vanilla SDXL are shown as references and excluded from this comparison.

Method	Base Model	Train-Free	CLIP-T↑	CLIP-I↑	DreamSim↓	Steps	Memory (GB)↓	Inference Time (s)↓
Vanilla SD1.5	-	-	0.8353	0.7474	0.5873	50	4.73	2.4657
Vanilla SDXL	-	-	0.9074	0.8165	0.5292	50	16.04	13.0890
BLIP-Diffusion	SD1.5	✗	0.7607	0.8863	0.2830	26	7.75	1.9284
Textual Inversion	SDXL	✗	0.8378	0.8229	0.4268	40	32.94	282.507
The Chosen One	SDXL	✗	0.7614	0.7831	0.4929	35	10.93	11.2073
PhotoMaker	SDXL	✗	0.8651	0.8465	0.3996	50	23.79	18.0259
IP-Adapter	SDXL	✗	0.8458	**0.9429**	**0.1462**	30	19.39	13.4594
ConsiStory	SDXL	✓	0.8769	0.8737	0.3188	50	34.55	34.5894
StoryDiffusion	SDXL	✓	0.8877	0.8755	0.3212	50	45.61	25.6928
Naive Prompt Reweighting (NPR)	SDXL	✓	0.8411	0.8916	0.2548	50	16.04	17.2413
1Prompt1Story (Ours)	SDXL	✓	**0.8942**	0.9117	0.1993	50	18.70	23.2088

Table 2: *User study* with 37 people to vote for the best consistent T2I generation method according to human preference.

Method	IP-Adapter	ConsiStory	StoryDiffusion	Ours
Percent (%)↑	8.60	13.00	29.80	**48.60**

Table 3: **Ablation study.** We evaluated the influence of each component in *1Prompt1Story*, including the *Singular-Value Reweighting* (SVR+ and SVR-), and *Identity-Preserving Cross-Attention* (IPCA).

Method	CLIP-T↑	CLIP-I↑	DreamSim↓
PCon; SVR+	0.8774	0.8886	0.2560
PCon; SVR-	0.8910	0.8904	0.2605
PCon; SVR+; SVR-	**0.8989**	0.8849	0.2538
PCon; SVR+; SVR-; IPCA (Ours)	0.8942	**0.9117**	**0.1993**

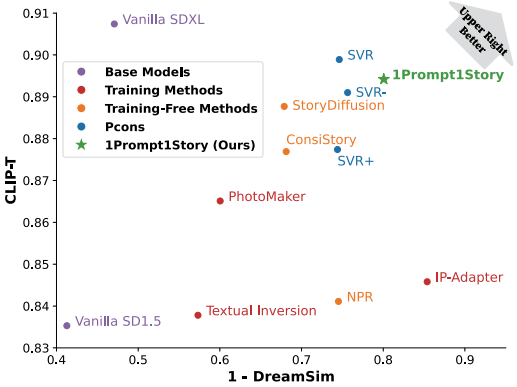

Figure 6: *Prompt alignment vs. identity consistency.* Our method *1Prompt1Story* is positioned in the upper right corner.

Quantitative Comparison. In Table 1, we illustrate the quantitative comparison with other approaches. In all evaluation metrics, *1Prompt1Story* ranks first among the training-free methods, and second when including training-required methods. Furthermore, compared to other training-free methods, our approach demonstrates a reasonable fast inference speed while achieving excellent performance. More specifically, our method *1Prompt1Story* achieves the CLIP-T score closely aligned with the vanilla SDXL model. In terms of identity similarity, measured by CLIP-I and DreamSim, our method ranks just below IP-Adapter. However, the high identity similarity of IP-Adapter mainly stems from its tendency to generate images with characters depicted in similar poses and layouts. To further explore this potential bias, we conducted a user study to investigate human preferences. Following ConsiStory, we also visualized our quantitative results using a chart, as shown in Fig. 6. Training-based methods, such as IP-Adapter and Textual Inversion, often overfit character identity and perform poorly on prompt alignment. In contrast, among training-free methods, our approach achieves the best balance in both prompt alignment and identity consistency.

User Study. In the user study, we compare our method with several state-of-the-art approaches, including IP-Adapter, ConsiStory, and StoryDiffusion. From our benchmark, we randomly selected 30 sets of prompts, each comprising four fixed-length prompts, to generate test images. Twenty participants were asked to select the image that best demonstrated overall performance in terms of identity consistency, prompt alignment, and image diversity. As shown in Table 2, the results indicated that our method *1Prompt1Story* aligns best with human preference. More details of the user study are shown in Appendix. F.

Ablation study. We performed an ablation study to analyze each component, as illustrated both qualitatively and quantitatively in Fig. 7 and Table 3. When using *Singular-Value Reweighting* exclusively with improving the express set as **SVR+** (that is, Eq. 2), the generated images blend with other descriptions, as can be seen in Fig. 7 (left, first row). Similarly, when *Singular-Value Reweighting* is only to weaken the suppress set as **SVR-** (i.e., Eq.3), the same issue appears in Fig. 7 (left, second row). In contrast, integrating both **SVR+** and **SVR-** in *Singular-Value Reweighting*

Figure 7: **Qualitative ablation study.** All ablated cases with incomplete components of *1Prompt1Story* struggle to achieve both prompt alignment and identity consistency as effectively as our full method.

Figure 8: **(Left):** Our method *1Prompt1Story* can integrate with ControlNet to enable spatial control for consistent character generation. **(Right):** Additionally, our method can also combine with other methods, such as PhotoMaker, to achieve real-image personalization with improved identity consistency.

can effectively mitigate blending in generated images (Fig. 7 (right, first row)). Although *Singular-Value Reweighting* can effectively resolve frame prompt blending issues, without *Identity-Preserving Cross-Attention*, there remains a weak inconsistency among the generated images. As shown in Fig. 7 (right, second row), the results indicate that using *Singular-Value Reweighting* and *Identity-Preserving Cross-Attention* achieves the best performance, as also evident in Table 3 (the last row). Additional results of ablation analysis and visualization are presented in the Appendix. C.

Additional applications. *1Prompt1Story* can also achieve spatial controls, integrating with existing control-based generative methods such as ControlNet (Zhang & Agrawala, 2023). As shown in Fig. 8 (left), our method effectively generates consistent characters with human pose control. Furthermore, our method can be combined with other approaches, such as PhotoMaker (Li et al., 2023b), to improve the consistency of identity with real images. By applying our method, the generated images more closely resemble the real identities, as demonstrated in Fig. 8 (right).

5 Conclusion

In this paper, we addressed the critical challenge of maintaining subject consistency in text-to-image (T2I) generation by leveraging the inherent property of *context consistency* found in natural language. Our proposed method, *One-Prompt-One-Story* (*1Prompt1Story*), effectively utilizes a single extended prompt to ensure consistent identity representation across diverse scenes. By integrating techniques such as *Singular-Value Reweighting* and *Identity-Preserving Cross-Attention*, our approach not only refines frame descriptions but also strengthens the consistency at the attention level. The experimental results on the *ConsiStory+* benchmark demonstrated the superiority of *1Prompt1Story* over state-of-the-art techniques, showcasing its potential for applications in animation, interactive storytelling, and video generation. Ultimately, our contributions highlight the importance of understanding context in T2I diffusion models, paving the way for more coherent and narrative-consistent visual output.

REFERENCES

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

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

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. *arXiv preprint arXiv:2311.10093*, 2023.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

Junhao Cheng, Xi Lu, Hanhui Li, Khun Loun Zai, Baiqiao Yin, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Autostudio: Crafting consistent subjects in multi-turn interactive image generation. *arXiv preprint arXiv:2406.01388*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023. doi: 10.1109/CVPR52729.2023.00276.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*, 2022.

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.

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. *International Conference on Learning Representations*, 2023a.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023b.

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*, pp. 1–10, 2023.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2862–2869, 2014.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Fx2SbBgcte.

Inhwa Han, Serin Yang, Taesung Kwon, and Jong Chul Ye. Highly personalized text embedding for image manipulation by stable diffusion. *arXiv preprint arXiv:2303.08767*, 2023a.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *Proceedings of the International Conference on Computer Vision*, 2023b.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002.

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*, pp. 8153–8163, 2024.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*, 2024.

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. VeRA: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NjNfLdxr3A.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024.

Senmao Li, Joost van de Weijer, Fahad Khan, Qibin Hou, Yaxing Wang, et al. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023a.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6329–6338, 2019.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023b.

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*, pp. 366–384. Springer, 2025.

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*, pp. 7817–7826, 2024.

Jinqi Luo, Kwan Ho Ryan Chan, Dimitris Dimos, and René Vidal. Knowledge pursuit prompting for zero-shot multimodal synthesis. *arXiv preprint arXiv:2311.17898*, 2023.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2427–2442, 2021.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pp. 70–87. Springer, 2022.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

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.

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*, pp. 8748–8763. PMLR, 2021.

Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2493–2502, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

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 (CVPR)*, pp. 10684–10695, 06 2022.

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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

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*, pp. 6527–6536, 2024.

RunDiffusion. Juggernaut x. In *RunDiffusion Tech Blog*, pp. 1, 2024.

Simo Ryu. Low-rank adaptation for fast text-to-image diffusion finetuning. https://github.com/cloneofsimo/lora, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Nikita Selin. Carvekit: Automated high-quality background removal framework. https://github.com/OPHoperHPO/image-background-remove-tool, 2023.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*, 2023.

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4733–4743, 2024.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. Storyimager: A unified and efficient framework for coherent story visualization and completion. *arXiv preprint arXiv:2404.05979*, 2024.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *arXiv preprint arXiv:2402.03286*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*, 2023.

Bingyuan Wang, Hengyu Meng, Zeyu Cai, Lanjiong Li, Yue Ma, Qifeng Chen, and Zeyu Wang. Magicscroll: Nontypical aspect-ratio image generation for visual storytelling via multi-layered semantic-aware denoising. *arXiv preprint arXiv:2312.10899*, 2023.

Qinghe Wang, Baolu Li, Xiaomin Li, Bing Cao, Liqian Ma, Huchuan Lu, and Xu Jia. Characterfactory: Sampling consistent characters with gans for diffusion models. *arXiv preprint arXiv:2404.15677*, 2024a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024b.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024a.

Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7685–7694, 2024b.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model, 2024. URL https://arxiv.org/abs/2407.08683.

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.

Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. Jedi: Joint-image diffusion models for finetuning-free personalized text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6786–6795, 2024.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024.

APPENDIX

A BOARDER IMPACTS AND LIMITATIONS

Boarder Impacts. The application of T2I models in consistent image generation offers extensive potential for various downstream applications, enabling the adaptation of images to different contexts. In particular, synthesizing consistent characters has diverse applications, however, it is a challenging task for diffusion models. Our *1Prompt1Story* can help the users customize their desired characters in different story scenarios, resulting in significant time and resource savings. Notably, current methods have inherent limitations, as discussed in this paper. However, our model can serve as an intermediary solution while offering valuable insights for further advancements.

Limitations. While our method *1Prompt1Story* can achieve high-fidelity consistent T2I generation, it is not free of limitations. Firstly, we have to know all the prompts in advance. Additionally, the length of the input prompt is constrained by the maximum capacity of the text encoder. Although we proposed a sliding window technique that facilitates infinite-length story generation in Appendix D.2, this approach may encounter issues where the identity of the generated images gradually diverges and becomes less consistent.

B IMPLEMENTATION DETAILS

B.1 MODEL CONFIGURATIONS

We generate subject-consistent images by modifying text embeddings and cross-attention modules at inference time, without any training or optimization processes. Our primary base model is the pre-trained Stable Diffusion XL (SDXL)[3]. SDXL has two text encoders: the CLIP L/14 encoder (Radford et al., 2021) and the OpenCLIP bigG/14 encoder (Cherti et al., 2023). We separately update the text embeddings produced by each encoder. For *Naive Prompt Reweighting*, we multiply the text embedding corresponding to the frame prompt that needs to be expressed by a factor of 2, while the text embedding corresponding to the *frame prompts* that need to be suppressed is multiplied by a factor of 0.5, keeping the $\boldsymbol{c}^{EOT}$ unchanged.

In our method, *1Prompt1Story*, we set the parameters as follows: $\alpha = 0.01, \beta = 1.5$ in Eq.2, and $\alpha' = 0.03, \beta' = 1.2$ in Eq.3. During the generation process, we initialize all frames with the same noise and apply a dropout rate of 0.5 to the token features in $\bar{\mathcal{K}}$ corresponding to $\mathcal{P}_0$. In the implementation of IPCA, the concatenated $\tilde{\mathcal{K}}$ and $\tilde{\mathcal{V}}$ are derived from the original text embeddings prior to applying SVR. We design an attention mask where all values in the column corresponding to $\mathcal{P}_i, i \in [1, N]$ are set to zero, while all other positions are set to one. The natural logarithm of this mask is then added to the original attention map. Our full algorithm is presented in Algorithm 1. Following (Tewel et al., 2024; Alaluf et al., 2024; Luo et al., 2023), we use Free-U (Si et al., 2024) to enhance the generation quality. All generated images based on SDXL are produced at a resolution of 1024×1024 using a Quadro RTX 3090 GPU with 24GB VRAM.

B.2 BENCHMARK DETAILS

To evaluate the effectiveness of our method, we developed *ConsiStory+*, an extended prompt benchmark based on ConsiStory (Tewel et al., 2024). We enhanced both the diversity and size of the original benchmark, which only comprised 100 sets of 5 prompts across 4 superclasses. Our expansion resulted in 200 sets, with each set containing between 5 and 10 prompts, categorized into 8 superclasses: humans, animals, fantasy, inanimate, fairy tales, nature, technology, and foods. The extended prompt benchmark was generated using ChatGPT 4.0-turbo[4], involving two main steps. First, we expanded the 100 prompt sets from the original benchmark, increasing each to a length of 5 to 10 prompts, as shown in Fig. 9 (left). Then, we generated new prompt sets for each of the new superclasses, as illustrated in Fig. 9 (right). The prompt sets collected through these two steps were combined to form our benchmark, *ConsiStory+*.

[3] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

[4] https://chatgpt.com/

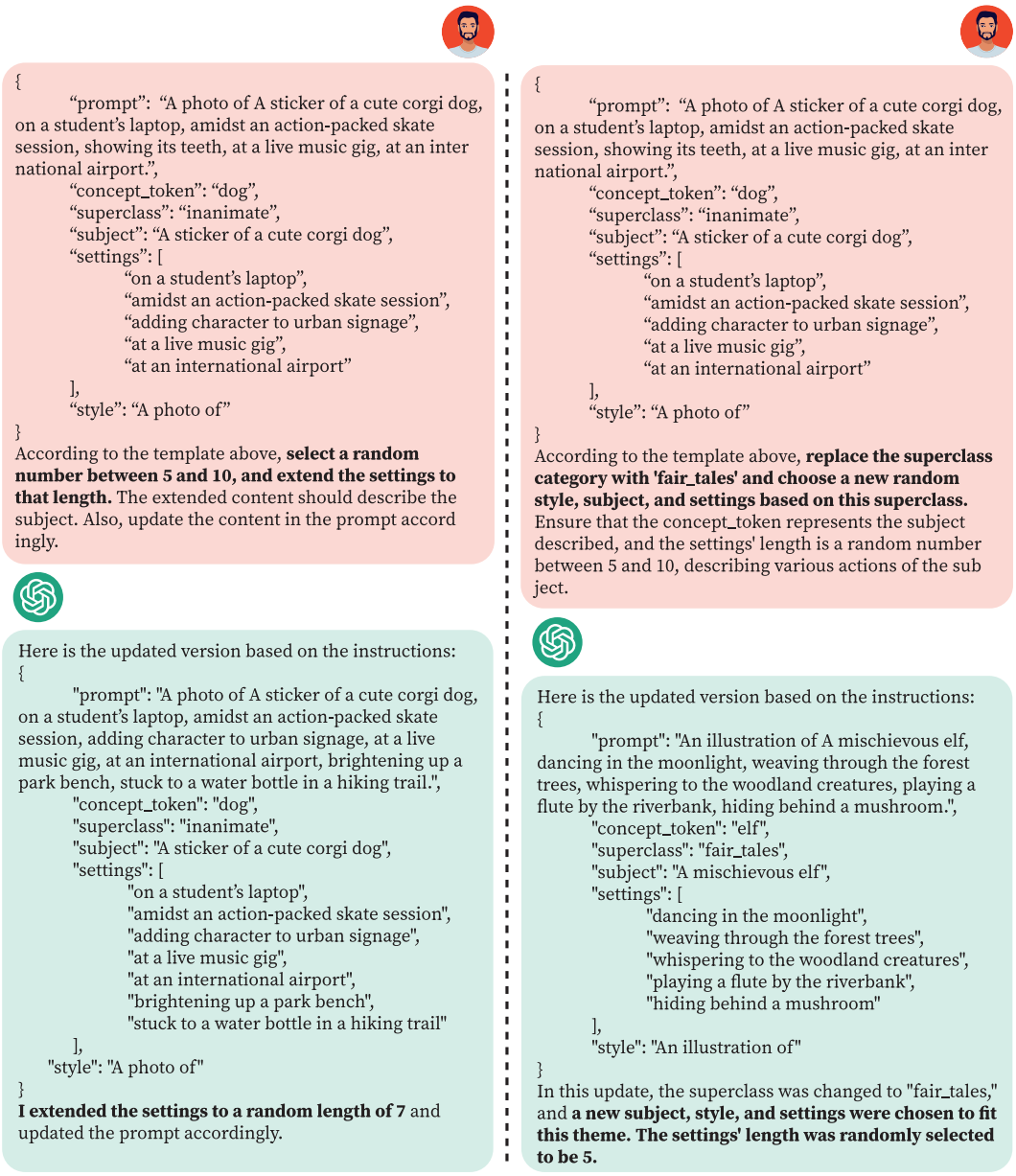

Figure 9: **(Left):** We expand the length of the original prompt sets to a random number between 5 and 10. **(Right):** We generate a new prompt set within one of the new superclass "fairy tales".

B.3 Comparison Method Implementations

We compare our method with all other approaches based on Stable Diffusion XL, except for BLIP-Diffusion (Li et al., 2024), which is based on Stable Diffusion v1.5[5]. The DDIM steps is set to the default value in the open-source code of each method. Below are the third-party packages we used for method implementations:

- The unofficial implementation of Textual Inversion (TI) (Gal et al., 2023a) at https://github.com/oss-roettger/XL-Textual-Inversion.
- The unofficial implementation of The Chosen One (Avrahami et al., 2023) at https://github.com/ZichengDuan/TheChosenOne.

[5]https://huggingface.co/runwayml/stable-diffusion-v1-5

Algorithm 1 *1Prompt1Story*

Input : A text embedding $\mathcal{C} = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \boldsymbol{c}^{\mathcal{P}_1}, \cdots, \boldsymbol{c}^{\mathcal{P}_N}, \boldsymbol{c}^{EOT}]$ and latent vector z_t.
Output: The subject consistency images $\mathcal{I}_1, \cdots, \mathcal{I}_N$.

for $j = 1, \ldots, N$ **do**
  // Singular-Value Reweighting
  $\hat{\mathcal{X}}^{exp} = [\hat{\boldsymbol{c}}^{\mathcal{P}_j}, \hat{\boldsymbol{c}}^{EOT}] \leftarrow \mathcal{X}^{exp} = [\boldsymbol{c}^{\mathcal{P}_j}, \boldsymbol{c}^{EOT}]$ (Eq. 2);
  for $k = [1, N] \setminus \{j\}$ **do**
    $\tilde{\mathcal{X}}^{sup} = [\tilde{\boldsymbol{c}}_k^{\mathcal{P}}, \tilde{\boldsymbol{c}}^{EOT}] \leftarrow [\boldsymbol{c}_k^{\mathcal{P}}, \hat{\boldsymbol{c}}^{EOT}]$ (Eq. 3);
  end
  $\tilde{\mathcal{C}} = [\boldsymbol{c}^{SOT}, \boldsymbol{c}^{\mathcal{P}_0}, \tilde{\boldsymbol{c}}^{\mathcal{P}_1}, \cdots, \hat{\boldsymbol{c}}^{\mathcal{P}_j}, \cdots, \tilde{\boldsymbol{c}}^{\mathcal{P}_N}, \tilde{\boldsymbol{c}}^{EOT}]$;

  // Identity-Preserving Cross-Attention
  for $t = T, \ldots, 1$ **do**
    $\tilde{\mathcal{K}}, \tilde{\mathcal{V}} \leftarrow \tilde{\mathcal{C}}$;
    $\bar{\mathcal{K}}, \bar{\mathcal{V}} \leftarrow \tilde{\mathcal{K}}, \tilde{\mathcal{V}}$;
    $\tilde{\mathcal{K}} = \texttt{Concat}(\tilde{\mathcal{K}}^\top, \bar{\mathcal{K}}^\top)^\top, \ \tilde{\mathcal{V}} = \texttt{Concat}(\tilde{\mathcal{V}}^\top, \bar{\mathcal{V}}^\top)^\top$;
    $\tilde{\mathcal{A}} \leftarrow \tilde{\mathcal{Q}}, \tilde{\mathcal{K}}$ (Eq. 4);
    $z_{t-1} \leftarrow \epsilon_\theta(z_t, t, \tilde{\mathcal{C}})$ with $\tilde{\mathcal{A}}$, $\tilde{\mathcal{V}}$;
  end
  $\mathcal{I}_j = D(z_0)$
end
Return $\mathcal{I}_1, \cdots, \mathcal{I}_N$.

- The official implementation of IP-Adapter (Ye et al., 2023) at https://github.com/tencent-ailab/IP-Adapter.
- The official implementation of PhotoMaker (Li et al., 2023b) at https://github.com/TencentARC/PhotoMaker.
- The official implementation of BLIP-Diffusion (Li et al., 2024) at https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion.
- The official implementation of StoryDiffusion (Zhou et al., 2024) at https://github.com/HVision-NKU/StoryDiffusion.

Since Consistory (Tewel et al., 2024) is not open-source, we reimplemented it ourselves. During the inference time, BLIP-Diffusion (Li et al., 2024), IP-Adapter (Ye et al., 2023), and PhotoMaker (Li et al., 2023b) all require a reference image as the additional input. To generate the reference image, we use their corresponding base models, providing the identity description as the input prompt. For example, if the full prompt is "a photo of a beautiful girl walking on the street", we use "a photo of a beautiful girl" to generate the reference image. The reference image is then used to generate all frames in the corresponding prompt set.

C Additional Ablation Study

C.1 Robustness to diverse description orders

To validate the robustness of our method regarding the order of *frame prompts*, we used the same three *frame prompts*: "wearing a scarf in a meadow", "playing in the snow", and "at the edge of a river" to create six different sequences for images generation. The *identity prompt* was consistently set to "a photo of a fox" and each sequence used the same seed for a generation. As shown in Fig. 10, our method *1Prompt1Story* generates images with identity consistency across different orders. Furthermore, the content of the images generated from varying sequences is closely aligned with the text descriptions, further demonstrating our method *Singular-Value Reweighting* effectiveness in suppressing content of unrelated *frame prompts*.

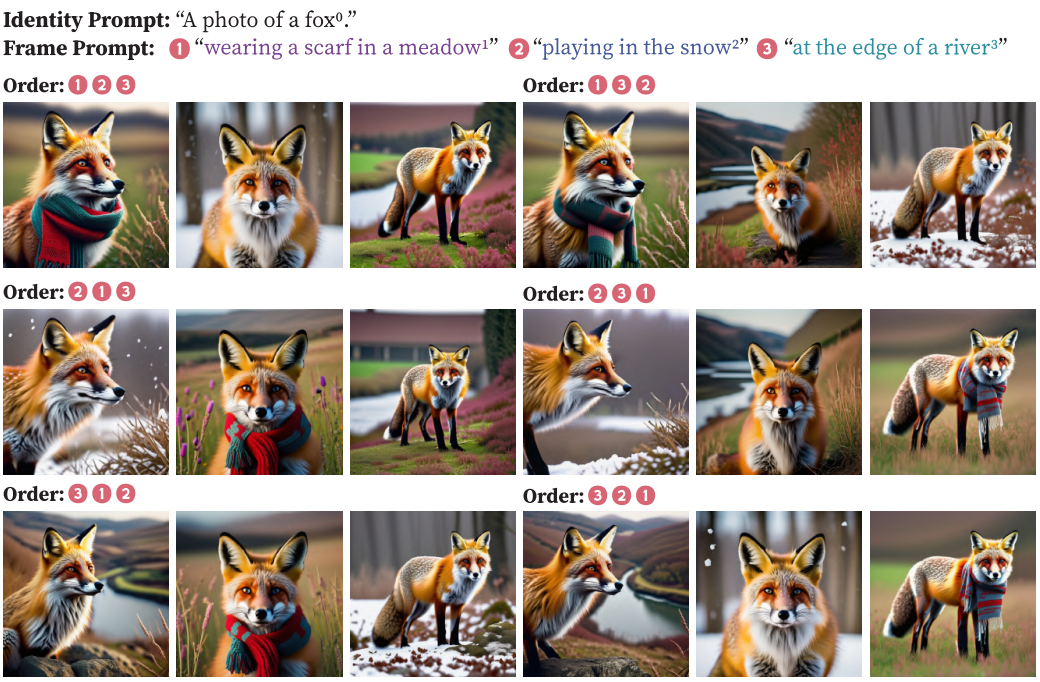

Figure 10: **Robustness to frame prompts order.** With the same set of *frame prompts* but in different orders, our method *1Prompt1Story* consistently generates images with a unified identity.

C.2 *Singular-Value Reweighting* ANALYSIS

Our *Singular-Value Reweighting* algorithm comprises two successive components: SVR+ enhances the *frame prompts* we wish to express, while SVR- iteratively weakens the *frame prompts* we aim to suppress. In our experiments, we first apply SVR+, followed by SVR-. In particular, we found that performing SVR- before SVR+ also yields similar results (see Fig. 11-left).

In the process of applying SVR-, we employed a strategy of iteratively suppressing each frame prompt. In fact, we could also concatenate the text embeddings corresponding to all frame prompts for suppression. To explore this, we conducted further ablation study specifically on the SVR- component. Assuming that we have n frames to generate, we discovered that merging the text embeddings corresponding to the $n-1$ frames we wish to suppress with $\boldsymbol{c}^{\text{EOT}}$ and subsequently performing the SVD decomposition does not effectively extract the main components of all *frame prompts* included in $\boldsymbol{c}^{\text{EOT}}$. Consequently, applying Eq. 3 to weaken the eigenvalues based on their magnitude fails to adequately eliminate the descriptions of all suppressed frames. we refer to this as "joint suppress", as illustrated in Fig. 11 (right, the first row). In contrast, if we handle each frame prompt to be suppressed individually and iteratively perform SVD and the operations from Eq. 3, which we term "iterative suppress", we can more effectively suppress all irrelevant *frame prompts*, as shown in Fig. 11 (right, the second row).

In our SVR, we enhance only the current frame prompt that needs to be expressed. An alternative option is to enhance the identity prompt simultaneously. We found that doing so can make the object's identity more consistent; however, it also introduces some negative effects, the background and subject's pose appearing more similar across images, as shown in Fig. 12. Furthermore, to demonstrate the role of the $\boldsymbol{c}^{\text{EOT}}$ in SVR, we conducted an ablation study on the $\boldsymbol{c}^{\text{EOT}}$ component. Specifically, we kept the $\boldsymbol{c}^{\text{EOT}}$ part of the text embedding unchanged during the SVR process and used this text embedding to generate images. As shown in Fig. 13, the results indicate that without performing SVR on the $\boldsymbol{c}^{\text{EOT}}$, the backgrounds of different frame prompts tend to blend together.

Figure 11: **(Left):** "SVR+ First" indicates that SVR+ is applied before SVR- in the *Singular-Value Reweighting* process, while "SVR- First" means the opposite order. We found that both sequences yield similar results (same seed). **(Right):** Compared to "Joint Suppress", "Iterative Suppress" is more effective at minimizing the influence of other *frame prompts* when generating images for the current frame. "Joint Suppress" produces images with similar backgrounds (the first row, first and third columns).

Figure 12: **SVR with identity enhancement.** The first row represents the original SVR with enhancements applied only to the frame prompt. The second row builds upon the original by further enhancing the identity prompt in the SVR+ module. The results indicate that while the second method improves identity consistency, it also leads to more similar object poses and backgrounds.

C.3 *Naive Prompt Reweighting* Ablation Study

Similar to the *Singular-Value Reweighting* (SVR) experiment, we conducted an ablation study to verify the effectiveness of *Naive Prompt Reweighting* (NPR) in terms of identity preservation and prompt alignment compared to our method *1Prompt1Story*. We denote NPR+ as applying a scaling factor of 2 to the text embedding corresponding to the current frame prompt that needs to be expressed. Conversely, NPR- denotes applying a scaling factor of 0.5 to the text embeddings of all other *frame prompts* that need to be suppressed. NPR represents the combination of both NPR+ and NPR- operations.

As shown in Fig. 14, images generated using the NPR+, NPR-, and NPR methods all exhibit varying degrees of interference from other *frame prompts*. In contrast, our method effectively removes irrelevant semantic information from other frame subject descriptions in the single-prompt setting, resulting in images that are more aligned with their corresponding *frame prompts*.

C.4 Seed Variety

Since our method *1Prompt1Story* does not modify the original parameters of the diffusion model, it preserves the inherent ability of the model to generate images with diverse identities and backgrounds using different seeds. By varying the initial noise while keeping the input prompt set constant, our method can produce a range of characters and backgrounds, all while maintaining strong identity consistency and prompt alignment, as shown in Fig. 15.

Figure 13: **Ablation study for c^{EOT}.** The left three images demonstrate the SVR process with a fixed c^{EOT}, while the right illustrates the SVR procedure described in the main text. The results indicate that keeping c^{EOT} unchanged leads to background blending across images generated for different frame prompts, highlighting the importance of updating c^{EOT} dynamically.

Figure 14: *Naive Prompt Reweighting* ablation study. NPR+, NPR-, and NPR are ineffective at suppressing the influence of other *frame prompts*. For example, the "puppy", which appears only in the frame prompt of the third frame, also shows up in the first and second frames using the aforementioned methods. In contrast, our method (the last row) effectively suppresses unwanted semantic information from other *frame prompts*.

D Additional results of our method *1Prompt1Story*

D.1 Consistent story generation with multiple subjects.

Our method is capable of generating stories involving multiple subjects. By specifying several subjects in the *identity prompt* and appending corresponding *frame prompts*, we can directly produce a series of images that maintain consistent identities across these subjects, as demonstrated in Fig. 16. However, this approach has a limitation: all generated images will include every character referenced in the *identity prompt*, which poses a constraint on the flexibility of our method.

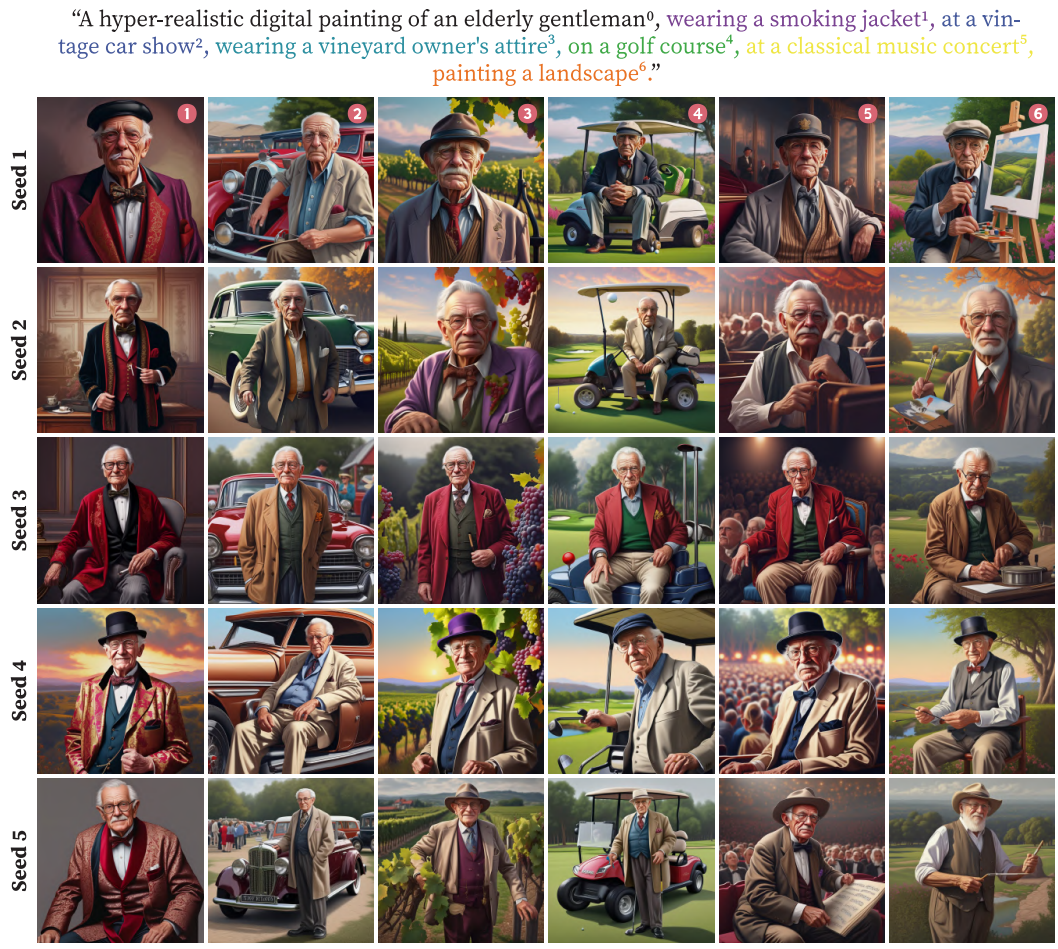

Figure 15: **Seed variation.** By using different seeds, our method *1Prompt1Story* can generate images with diverse backgrounds while maintaining a consistent identity.

D.2 Story generation of any length.

To generate stories of any length, we designed a "sliding window" technique to overcome the input text length limitations of diffusion models like SDXL. Suppose we aim to generate a story with n images, each corresponding to n *frame prompts*, using a window size t, where $t < n$. Similarly, we represent the *identity prompt* as $\mathcal{P}_0$ and the *frame prompts* as $\mathcal{P}_i$, where $i \in [1, n]$. For generating the image corresponding to the i-th frame, if $i \leq t$, we use $\mathcal{P} = [\mathcal{P}_0; \mathcal{P}_1; \ldots; \mathcal{P}_t]$ as input prompt and apply our method *1Prompt1Story* to generate the images. If $i > t$, we use $\mathcal{P} = [\mathcal{P}_0; \mathcal{P}_{i-t+1}; \ldots; \mathcal{P}_i]$ to generate the images. As shown in Fig. 19, we applied an ultra-long prompt to generate 42 images with consistent identities, using a window size of 10.

D.3 Combine with different diffusion models.

Since our method exclusively modifies the text-embedding and cross-attention modules of the diffusion model, it can be directly adapted to other diffusion models. In this study, we implemented our approach within the SDXL framework. Other models utilizing the SDXL framework, such as playground-v2.5[6], RealVisXL_V4.0[7] and Juggernaut-X-v10[8], can apply our method without any additional modifications or fine-tuning. Our experimental results (see Fig. 20) indicate that these

[6] https://huggingface.co/playgroundai/playground-v2.5-1024px-aesthetic
[7] https://huggingface.co/SG161222/RealVisXL_V4.0
[8] https://huggingface.co/RunDiffusion/Juggernaut-X-v10

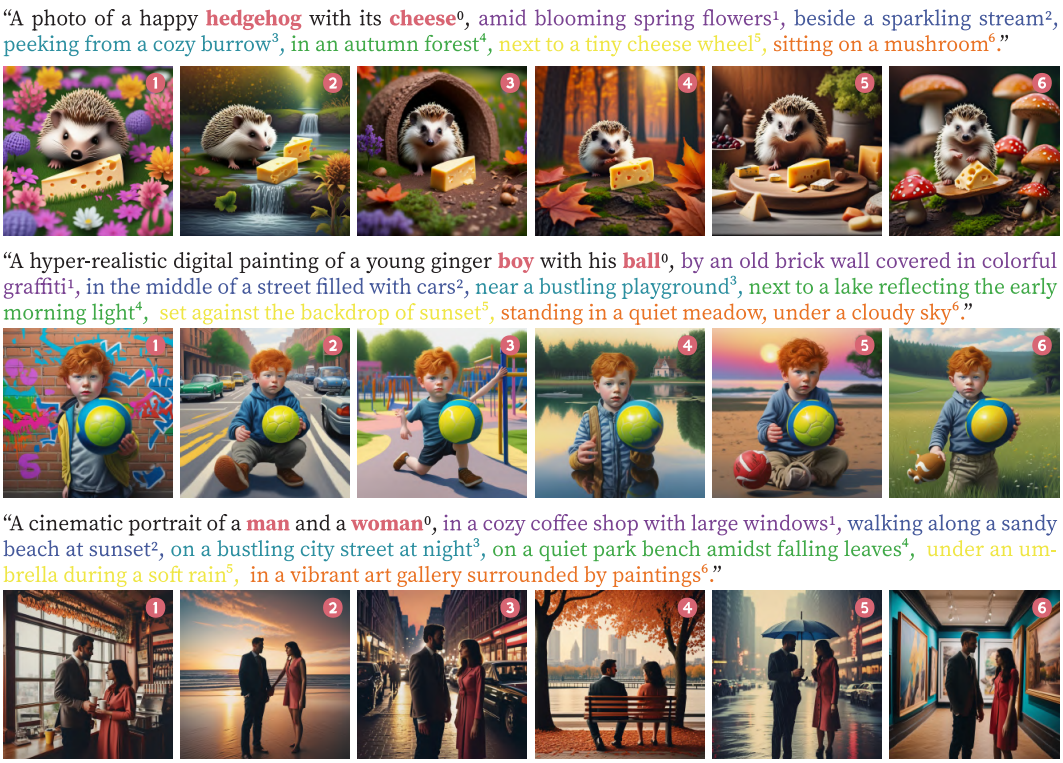

Figure 16: **Multi-subject story generation.** By defining multiple subjects in the *identity prompt*, our method generates images featuring multiple characters, each maintaining good identity consistency.

Figure 17: **Additional result with PhotoMaker.** We compared additional results of our method combined with PhotoMaker, where a lower DreamSim score indicates better ID consistency between the generated images. The results demonstrate that our method has the potential to enhance the performance of PhotoMaker.

models can also achieve image generation with enhanced identity consistency when employing our method *1Prompt1Story*.

E Additional Experiments

E.1 Additional Prompt Alignment Metrics

In addition to the primary evaluation metrics, we conduct an experiment using the recent prompt alignment metrics DSG(Cho et al., 2023) and VQAScore(Lin et al., 2025). Both DSG and VQA are metrics that measure the consistency between images and text by evaluating questions and their corresponding answers. These metrics have been shown to provide more reliable strengths in fine-

Metric	SD1.5	SDXL	BLIP-Diffusion	Textual Inversion	The Chosen One	PhotoMaker	IP-Adapter	ConsiStory	Story Diffusion	NPR	Ours
VQAScore↑	0.7157	0.8473	0.5735	0.6655	0.6990	0.8178	0.7834	0.8184	**0.8335**	0.8044	<u>0.8275</u>
DSG w/ dependency↑	0.7354	0.8524	0.6128	0.7219	0.6667	0.8108	0.7564	0.8196	0.8400	<u>0.8407</u>	**0.8520**
DSG w/o dependency↑	0.8095	0.8961	0.6909	0.8051	0.7495	0.8700	0.8122	0.8696	0.8853	<u>0.8863</u>	**0.8945**
FID↓	-	-	65.32	48.94	83.74	55.27	66.76	45.20	51.63	**44.02**	<u>44.16</u>

Table 4: **Additional metircs comparison.** SD1.5 and SDXL are shown as references and excluded from this comparison. The **bold** and <u>underlined</u> are the best and second best results respectively.

grained diagnosis and align closely with human judgment. We present our comparison with all other methods in Table 4, results show that our method *1Prompt1Story* outperforms other training-based methods and achieves the highest value on the DSG metric.

E.2 Visual Quality Comparsion

To evaluate the impact of different methods on image quality under ID consistency generation, we use images generated by the base model as the real dataset and images generated by each method itself as the fake dataset. Then, we calculate the FID(Heusel et al., 2017). As shown in Table 4 (the last row), *Naive Prompt Reweighting* (NPR) and our method *1Prompt1Story* achieved the best and second-best results in terms of FID. This indicates that our method has a smaller impact on the image generation quality of the base model compared to other methods.

E.3 Context Consistency in Text Embeddings

Besides the separate t-SNE dimensionality reduction conducted for multi-prompt and single-prompt setups in sec. 3.1.1, we extended our analysis by performing a joint t-SNE reduction on the combined text embeddings from both setups. This unified approach allows for a direct visual comparison of the embeddings' spatial arrangements within the text representation space. As illustrated in Fig. 18 (left), the text embeddings originating from the multi-prompt setup remain widely dispersed (red dots), indicative of their diverse semantic properties. Conversely, embeddings from the single-prompt setup (blue dots) exhibit noticeably tighter clustering. To substantiate these observations, we also perform statistical analysis on our benchmark dataset, as shown in Fig. 18 (right).

F User Study Details

In the user study, we compared our method with three state-of-the-art approaches: IP-Adapter, Consistory, and Story Diffusion. We selected 30 prompt sets from our *ConsiStory+* benchmark to generate test images, with each prompt set producing four frames.

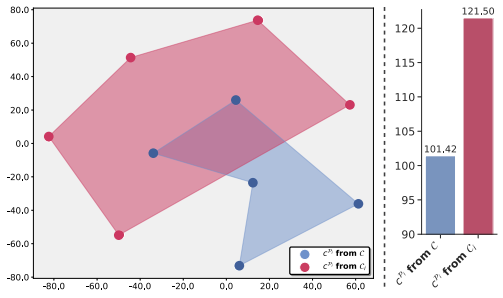

Figure 18: **Additonal t-SNE visualization of text embeddings (Left) and statistical results (Right).**

In the questionnaire, participants were first provided with guidance on selecting images. They were instructed to choose the set that exhibited the most balanced performance across three criteria: identity consistency, prompt alignment, and image diversity, according to their personal preferences. As illustrated in Fig. 21, we detailed these criteria at the beginning of the questionnaire. Additionally, we provided an example to demonstrate our recommended best choice, including justifications for both selecting and not selecting each set, thereby aiding participants in making informed decisions.

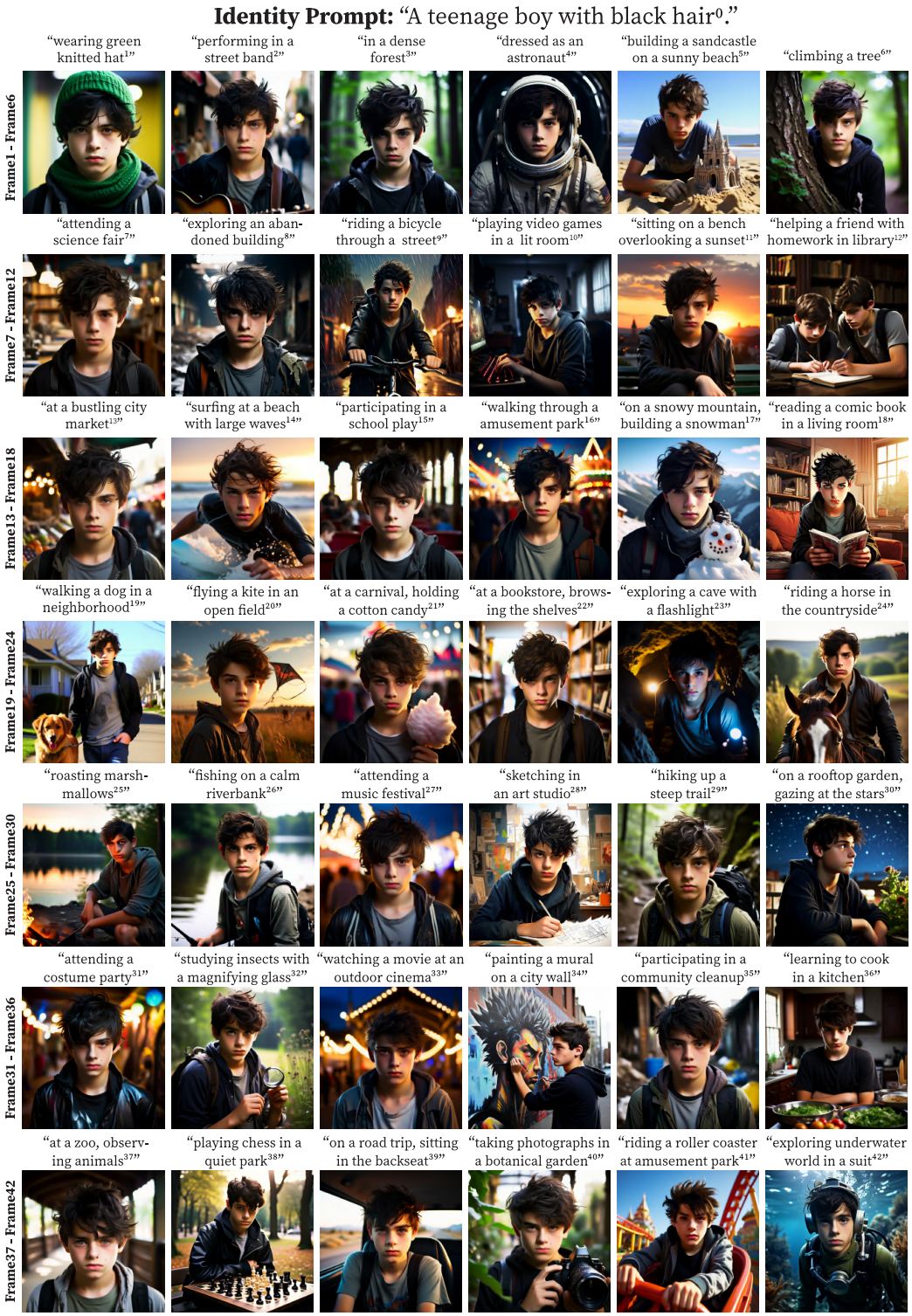

Figure 19: **Long story generation.** By using the "sliding window" technique, our method *1Prompt1Story* can generate stories of any length with consistent identity throughout.

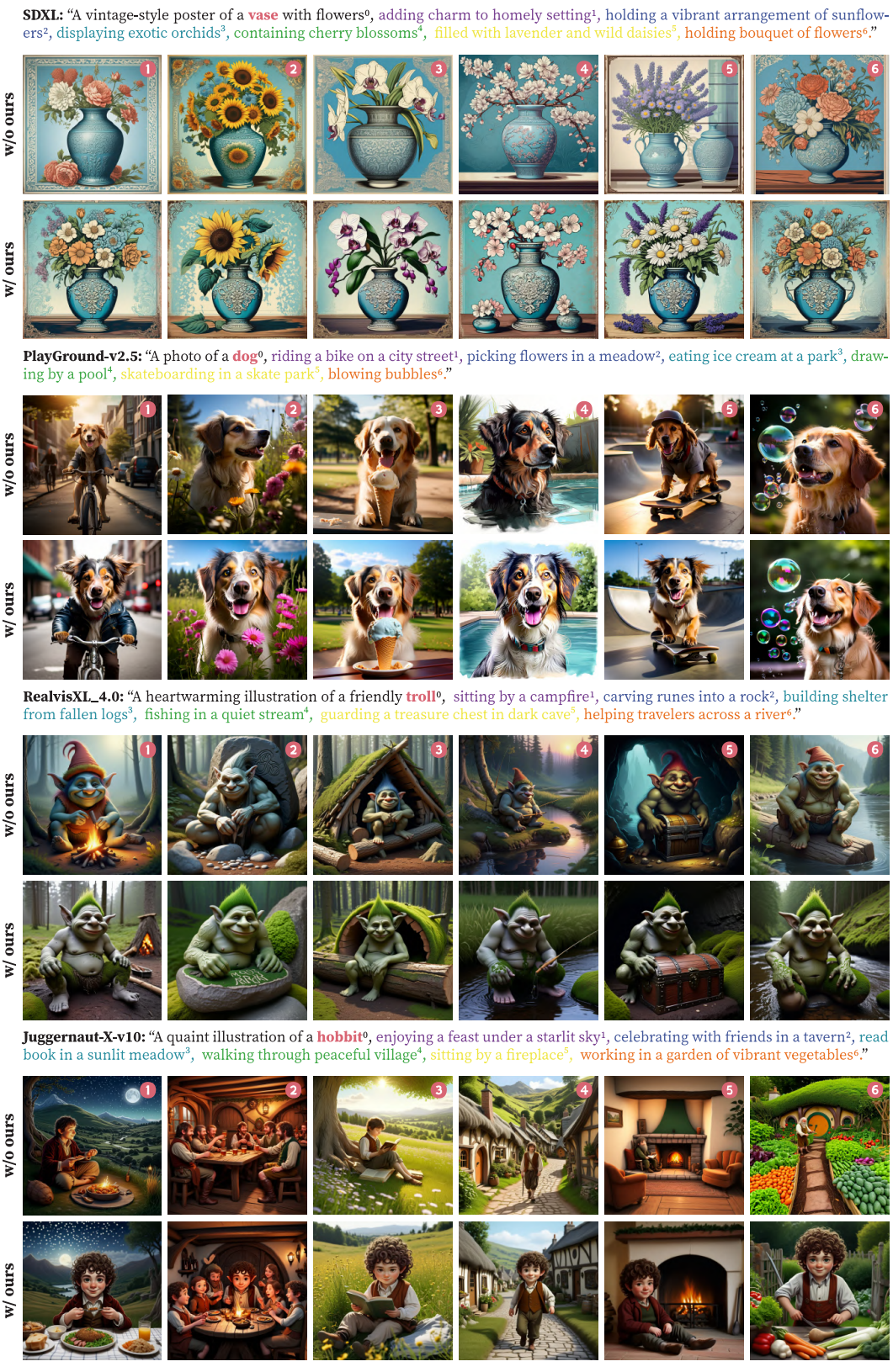

Figure 20: **Evaluation with different models.** We test our method on various T2I diffusion models, and without requiring fine-tuning, our approach could directly generate images with a consistent identity.

Selection Guidance

In this survey, you will evaluate four sets of images based on three criteria: **"Identity Consistency" "Prompt Alignment"** and **"Image Diversity"**. Your task is to select the set that performs best across all three aspects.

Identity Consistency: Refers to the visual coherence of the subject's appearance across the set, indicating that the same subject is depicted in all images.

Prompt Alignment: Indicates how well each image in the set matches its corresponding text description.

Image Diversity: Refers to the variety of poses, object arrangements, and overall composition within the set of images.

Example

Each row represents one of the four image sets: A, B, C, and D. Each column corresponds to the same frame descriptions:
['wearing a superhero cape', 'at the beach', 'wearing a headscarf', 'wearing a birthday hat'].

In this example, the best choice is set A (the first row).

Reason for Selection

Set A (the first row) performs well in terms of "Identity Consistency," "Text Alignment," and "Image Diversity".
Set B (the second row) is not chosen because its identity consistency is poor.
Set C (the third row) is not selected despite its high identity consistency because its text alignment and image diversity are lacking.
Set D (the fourth row) is also not chosen due to its poor identity consistency.

Figure 21: **User study questionnaire.** Before filling out the questionnaire, participants were provided with selection guidelines, including detailed explanations of the three evaluation criteria: identity consistency, prompt alignment, and image diversity. Additionally, an example was provided, along with our recommended best choice and the reasoning behind the selection.

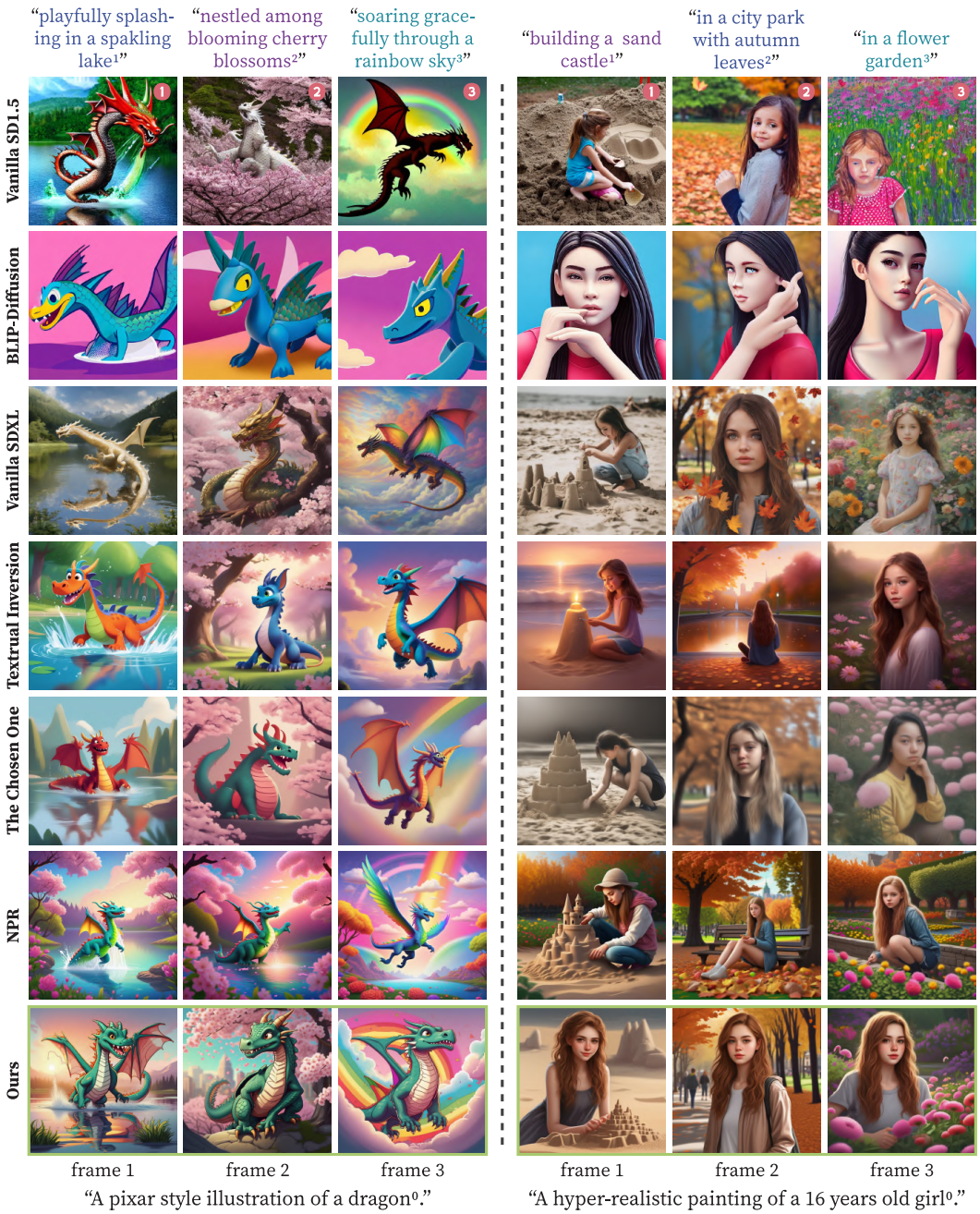

Figure 22: **Additional qualitative comparison.** We also compared our method with other existing approaches. The characters generated by vanilla SD1.5 and vanilla SDXL exhibit significant variations in both form and appearance. In contrast, some training-based methods, such as Textual Inversion and The Chosen One, generate characters with consistent forms, but their appearance lacks similarity. While NPR can produce characters with consistent identities, the backgrounds often blend across images. In contrast, our method not only ensures identity consistency but also generates backgrounds that closely align with the corresponding text descriptions.