STORYCTRL: CUSTOMIZED STORY VISUALIZATION WITH FINE-GRAINED CONTROL

Anonymous authors
Paper under double-blind review

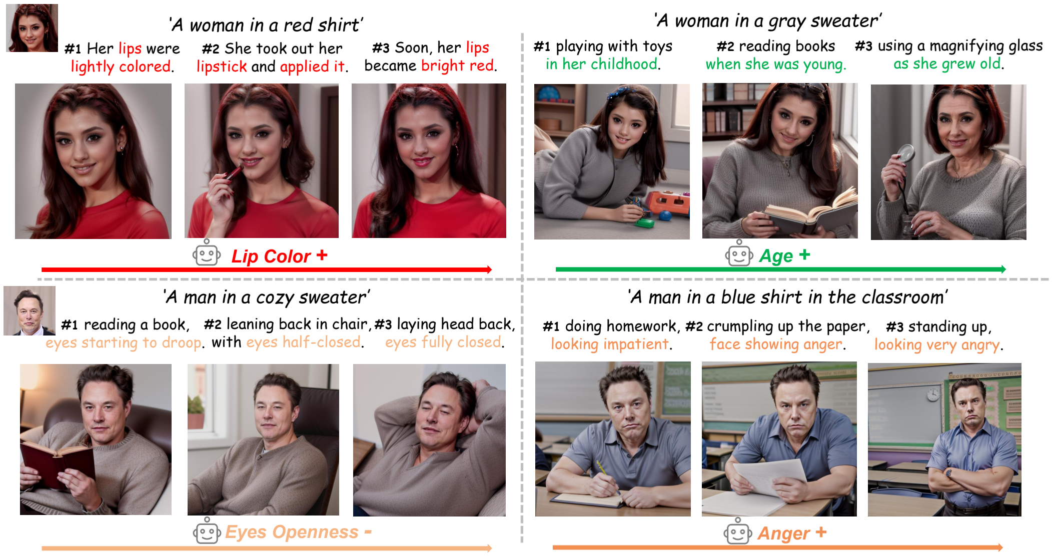

'A woman in a red shirt'
#1 Her lips were lightly colored.
#2 She took out her lipstick and applied it.
#3 Soon, her lips became bright red.
Lip Color +
'A woman in a gray sweater'
#1 playing with toys in her childhood.
#2 reading books when she was young.
#3 using a magnifying glass as she grew old.
Age +
'A man in a cozy sweater'
#1 reading a book, eyes starting to droop.
#2 leaning back in chair, with eyes half-closed.
#3 laying head back, eyes fully closed.
Eyes Openness -
'A man in a blue shirt in the classroom'
#1 doing homework, looking impatient.
#2 crumpling up the paper, face showing anger.
#3 standing up, looking very angry.
Anger +

Figure 1: Our model StoryCtrl allows users to generate visual stories that not only preserve identity fidelity and ensure inter-frame consistency, but also enable fine-grained control over the specific attributes of characters in generated stories.

ABSTRACT

Recent advancements in story visualization have achieved significant progress through text-to-image (T2I) models that generate coherent image sequences aligned with narratives. However, despite these advancements, generating customized story visualizations remains challenging. Current methods primarily address identity (ID) fidelity and consistency across frames but overlook the fine-grained control of character attributes, leading to suboptimal generation results. To tackle these limitations, we propose StoryCtrl, an innovative framework that not only preserves identity fidelity but also enables fine-grained control over specific character attributes in generated stories. The proposed framework consists of four key components. First, the CtrlGAN Encoder extracts ID information and inverts visual features to the W+ latent space. Second, the Story-aware Module (SaM) captures attribute changes within the narrative context and assists CtrlGAN in making adjustments during the image encoding stage, enabling fine-grained attribute control. Third, we introduce ID-Consis Attention (ICA), which ensures consistency in generated story sequences. Finally, we incorporate Customized Guidance Fusion (CGF), which integrates reference image features and prompts to enhance customization. To the best of our knowledge, we are the first to introduce an expanded definition of story visualization and present a method for generating fine-grained character attributes. Extensive experiments demonstrate that our method achieves state-of-the-art (SOTA) performance in customized story visualization.

1 Introduction

Customized Story Visualization (CSV) is designed to enhance narrative engagement by generating visually consistent stories that adhere to the identities of customized characters. Combined with diffusion models, recent customized story visualization methods demonstrate wide applications Yin et al. (2022) in comic creation and photo blog generation. Early works on customized generation Ye et al. (2023); Wang et al. (2024); Li et al. (2024b), primarily focus on high ID fidelity but fail to address the consistency of the generated character during story visualization. Subsequent works, including StoryDiffusion Zhou et al. (2024a), TaleCrafter Gong et al. (2023), and StoryMaker Zhou et al. (2024b), aim to solve this problem and improve consistency across generated images. Despite their effectiveness, these methods all neglect subtle character attribute changes in the story text, failing to generate character visuals that align with textual descriptions. We refer to this as a lack of fine-grained control over character attributes based on the text.

To address the aforementioned problems, we propose StoryCtrl, an effective framework that maintains ID fidelity and enables fine-grained control over the specific attributes of characters in generated stories. Our StoryCtrlconsists of several key components: CtrlGAN Encoder, Story-aware Module (SaM), Customized Guidance Fusion (CGF) and ID-Consis Attention (ICA).

Given a reference image of customized identity, we first use the CtrlGAN Encoder to extract ID information. While maintaining identity fidelity, our CtrlGAN disentangles ID-irrelevant facial attributes, enabling fine-grained control based on the text. To understand the variation of specific attributes in the story description, we use the Story-aware Module. It captures text trends in the story, enabling text-aligned fine-grained control and assisting CtrlGAN in making adjustments during the image encoding stage. We further enhance the integration of customized guidance into the generation process through CGF in a residual manner, which achieves a balance between the text prompts and customized guidance. Moreover, to ensure inter-frame consistency in the generated story, we introduce ID-Consis Attention, which establishes interactions across frames of the story to maintain the identity consistency of our customized character.

Through comprehensive qualitative and quantitative comparisons, we have demonstrated both the effectiveness and superiority of StoryCtrl. Furthermore, we also conducted extensive ablation studies to validate the effectiveness of each component of our method. Our contributions can be summarized as follows:

- We are the first to observe that existing story visualization methods primarily focus on ID preservation, yet often fail to generate visual content that accurately reflects the narrative text, making it difficult to satisfy human expectations. To address this limitation, we expand the task definition by introducing fine-grained control over character attributes, enabling more semantically aligned and higher-quality visual storytelling.
- To this end, we propose StoryCtrl, a novel framework composed of four key components. The Story-aware Module (SaM) enhances the textual understanding of the story prompt and facilitates attribute conditioning. It works alongside CtrlGAN, which extracts comprehensive character representations to preserve ID fidelity while effectively disentangling ID-irrelevant attributes. Next, we incorporate Customized Guidance Fusion (CGF) into the T2I generation pipeline, enabling controllable visual synthesis. Furthermore, to ensure coherence across the entire image sequence, we introduce ID-Consis Attention, which establishes consistency across generated frames.
- Comprehensive qualitative and quantitative experiments, along with human preference studies, demonstrate that StoryCtrl outperforms existing state-of-the-art methods in the task of customized story visualization.

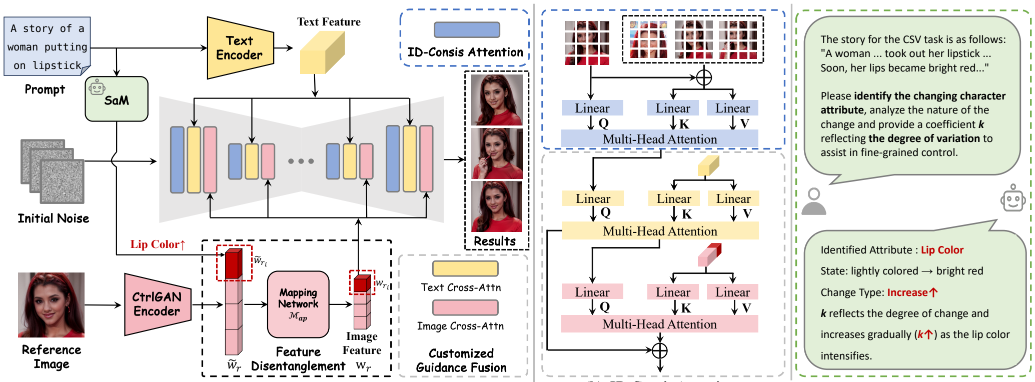
2 Related Work

Customized Image Generation. Owing to the powerful generative capabilities of the diffusion models, customized image generation has achieved remarkable progress. Early approaches such as DreamBooth Ruiz et al. (2023) and Textual Inversion Gal et al. (2022) fine-tune the diffusion model for the target concept by using a limited set of subject images. However, these methods are plagued by time-consuming issues, as the slow optimization process prior to inference demands substantial computational resources and time. Recent studies Chen et al. (2024a;b); Jia et al. (2023); Shi et al. (2024); Wei et al. (2023) aim to perform customized generation with a single image through a single forward pass, which has significantly expedited the customization process. ID-Preservation, aimed at generating images with a specified ID, is a prominent area in customized image generation.

Recent works exemplified by IPAdapter Ye et al. (2023) and PhotoMaker Li et al. (2024b) leverage models that have undergone pre-training on large datasets to preserve facial features, making them highly effective in personalized generation. However, we observe that these approaches not only cause the synthesized faces to overfit to the reference, sharing too many similar attributes, but also fail to provide fine-grained control over these attributes.

Story Visualization. Story visualization Liu et al. (2024); Mao et al. (2024) is tasked with the generation of images featuring consistent content, namely maintaining consistent content across a sequence of generated images. StoryDiffusion Zhou et al. (2024a) utilizes a consistent self-attention mechanism. This mechanism adapts information from other images within the batch to guarantee character consistency throughout the storytelling sequences. In contrast to StoryDiffusion, which draws adaptation from full-scale images, ConsiStory Tewel et al. (2024) adopts a subject-driven shared attention block. This block solely adapts information from masked subjects, and furthermore, correspondence-based feature injection is implemented to enhance the subject consistency between images. StoryMaker Zhou et al. (2024b), on the other hand, differs from the above methods by concentrating on generating images with consistent faces when given references. It is worth noticing that our proposed StoryCtrl, with merely a single reference image, not only preserves the facial fidelity of characters but also attains fine-grained control while maintaining inter-frame consistency. This is fundamental in customized story visualization, enhancing the overall quality of visualized stories.

3 Methodology



The story for the CSV task is as follows: "A woman ... took out her lipstick ... Soon, her lips became bright red..."

Please **identify the changing character attribute**, analyze the nature of the change and provide a coefficient ***k*** reflecting **the degree of variation** to assist in fine-grained control.

Identified Attribute : **Lip Color**

State: lightly colored → bright red

Change Type: **Increase↑**

k reflects the degree of change and increases gradually (***k*↑**) as the lip color intensifies.

(a) Overall Pipeline of StoryCtrl

(b) ID-Consis Attention &Customized Guidance Fusion

(c) Story-aware Module (SaM)

Figure 2: **Overviews of StoryCtrl.** (a) Overall architecture of our method, which is designed to generate an inter-frame consistent visual story with fine-grained control based on reference images and prompts. (b) Newly integrated attention layers of StoryCtrl. ID-Consis Attention (ICA) facilitates cross-image interactions within batches to maintain ID consistency, while Customized Guidance Fusion (CGF) intricately merges fine-grained ID information into the generation process. (c) Story-aware module (SaM), harnesses capabilities from large language models (LLM) to interpret text trends in narratives, enabling precise adjustments during the image encoding phase for fine-grained control.

Given a reference image of a customized identity (ID) and a story in plain text, the purpose of Customized Story Visualization (CSV) is to generate a set of images depicting consistent characters in different scenarios. In this section, we will introduce the methodology of StoryCtrl in detail, which mainly includes several components: CtrlGAN Encoder, Story-aware Module (SaM), Customized Guidance Fusion (CGF), and ID-Consis Attention (ICA). The pipeline is illustrated in Fig. 2.

3.1 Preliminary

DDPM. Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020) are generative models that reconstruct a predefined forward Markov chain $x_0, \ldots, x_T$. Given data $x_0 \sim q(x_0)$, the forward process follows a Gaussian transition:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathrm{I}), \quad (1)$$

where $\beta_t \in (0, 1)$ is a variance schedule. Starting from a prior $p(x_T) = \mathcal{N}(x_T; 0, \mathbb{I})$, a reverse process predicts $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ using a denoising network ϵ_θ:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, \boldsymbol{\tau}, t), \Sigma_\theta(\boldsymbol{x}_t, \boldsymbol{\tau}, t)), \tag{2}$$

where $\boldsymbol{\tau}$ is the textual prompt, and μ_θ, Σ_θ are predicted by ϵ_θ.
Although DDPM operates in pixel space, it is computationally intensive. Stable Diffusion Rombach et al. (2022); Podell et al. (2023) mitigates this by performing diffusion in the latent space of a variational autoencoder (VAE) Kingma & Welling (2013), significantly improving efficiency.

StyleGAN Latent Space. GANs Liu et al. (2022) have gained significant attention for their well-structured and highly interpretable latent space. Works like PGGAN Karras et al. (2018) and StyleGAN Karras (2019) explore the interpretable semantics inside the latent space of fixed GAN models and turn unconstrained GANs to controllable GANs by varying the latent code. Image2StyleGAN Shen et al. (2020) addresses the challenge of embedding a given image into StyleGAN's latent space. This embedding enables semantic image editing of existing images. Inspired by the disentanglement and editability of StyleGAN's latent space, we aim to introduce it for ID extraction while disentangling ID-irrelevant attributes to enable fine-grained control.

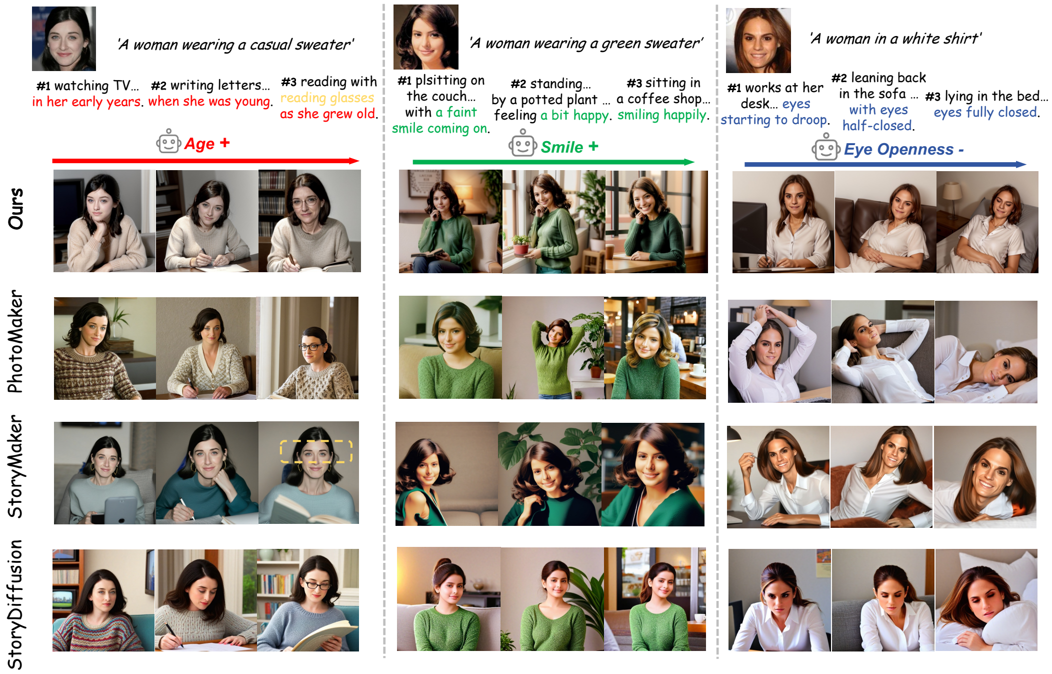

'A woman wearing a casual sweater'
#1 watching TV... in her early years.
#2 writing letters... when she was young.
#3 reading with reading glasses as she grew old.
Age +
'A woman wearing a green sweater'
#1 plsitting on the couch... with a faint smile coming on.
#2 standing... by a potted plant ... feeling a bit happy.
#3 sitting in a coffee shop... smiling happily.
Smile +
'A woman in a white shirt'
#1 works at her desk... eyes starting to droop.
#2 leaning back in the sofa ... with eyes half-closed.
#3 lying in the bed... eyes fully closed.
Eye Openness -
Ours
PhotoMaker
StoryMaker
StoryDiffusion

Figure 3: Qualitative comparisons between our StoryCtrland other methods regarding fine-grained control over character attributes (Age, Smile).

3.2 ID Preservation with Fine-Grained Control

High-quality Customized Story Visualization (CSV) must satisfy two fundamental requirements:
ID Preservation: The primary objective is to ensure that the characters in the generated frame sequence maintain the unique ID information derived from the reference image.
Accurate Visual Representation: Secondary, the character visuals must visually align with the narrative description, necessitating fine-grained control over character attributes based on the text.
To address these requirements, we propose an effective methodology for extracting ID information from the reference image, seamlessly integrating it into the generative model, and simultaneously enabling fine-grained control over character attributes.

3.2.1 CtrlGAN Encoder.

Previous approaches to customized image generation Ruiz et al. (2023); Xiao et al. (2024); Ye et al. (2023) primarily relied on a pre-trained CLIP image encoder Radford et al. (2021a) to extract image features from the reference image. While these methods have proven capable of maintaining identity, they are often constrained by inherent trade-offs. Specifically, they struggle to preserve identity while generating varied facial features (e.g., makeup, expression, etc.), meaning that they fail to achieve fine-grained control that aligns with the text description.

Building on previous findings regarding intrinsic ID features Yuan et al. (2024), our key insight is to disentangle the various semantics of facial attributes, allowing fine-grained attribute control. Thanks to analysis Shen et al. (2020); Härkönen et al. (2020) on how different semantics are encoded in the latent space, we are motivated to propose CtrlGAN Encoder as an image encoder to extract ID information from the reference image. Specifically, the CtrlGAN Encoder inverts the visual concept of a customized identity into W_+ latent space Shen et al. (2020) of StyleGAN. Given a reference image I_r, we denote CtrlGAN as our image encoder $\mathcal{C}_{\text{trl}}$. The corresponding W_+ space vector $\tilde{w}_r$ is obtained through:

$$\tilde{w}_r = \mathcal{C}_{\text{trl}}(I_r). \tag{3}$$

3.2.2 Story-aware Module.

We choose the CtrlGAN Encoder to extract ID information from the reference image, ensuring ID preservation. By disentangling ID-irrelevant facial attributes, we can now achieve fine-grained control based on the textual description. Specifically, we propose the Story-aware Module, which captures text trends in the story through the Large Language Model (LLM), further assists CtrlGAN in making adjustments during the image encoding stage to enable fine-grained control.

Disentangling different semantic spaces, $i \in \{0, 1, \ldots, N\}$ represents the identity-irrelevant attributes, which are used for fine-grained control. Given a story for visualization, our Story-aware Module can identify the gradual variation in the specific attribute, denoted as Δw direction of the i attribute derived from the W_+ space.

$$\tilde{w}_{r_i} = \tilde{w}_{r_i} + \Delta w \cdot k, \quad i \in \{0, 1, \ldots, N\}, \tag{4}$$

where $\tilde{w}_{r_i}$ represents a subspace of $\tilde{w}_r$ with respect to the specific attribute i, the attribute direction Δw can be obtained following Shen et al. (2020), k can be either positive or negative, depending on how the specific attribute changes according to the text description. For example, given a story about a woman applying lipstick, our Story-aware Module can identify the gradual increase in lipstick attribute, where the gradual growth of k leads to the gradual darkening of the lipstick color, thus enabling fine-grained control. In Fig. 2, we provide a simplified template that includes our instructions alongside relevant in-context examples.

3.2.3 Mapping Network.

Since the original W_+ space is designated for StyleGAN generation, we introduce a mapping network $\mathcal{M}_{\text{ap}}$ following research on editing in the StyleGAN space Karras (2019); Shen et al. (2020); Li et al. (2024a); Härkönen et al. (2020); Karras et al. (2018), which projects the vector $\tilde{w}_r$ into the dimension aligned with Stable Diffusion, denoted as w_r.

$$w_r = \mathcal{M}_{\text{ap}}(\tilde{w}_r). \tag{5}$$

3.2.4 Customized Guidance Fusion.

To further enhance the integration of fine-grained ID information from w_r into the generation process, we introduce the Customized Guidance Fusion strategy. This approach processes two inputs: the output image feature f_s from the text cross attention and the character embedding w_r. Through the decoupled cross-attention layer Ye et al. (2023); Wei et al. (2023), it produces a character representation f_g that incorporates detailed control signals. Given the projection matrices W_q^g, W_k^g and W_v^g of the cross-attention layer, we first obtain the corresponding vectors:

$$Q_g = f_s W_q^g, \quad K_g = w_r W_k^g, \quad V_g = w_r W_v^g. \tag{6}$$

Next, we fuse the image features and customized guidance through cross-attention operations:

$$f_g = \text{Softmax}\left(\frac{Q_g K_g^T}{\sqrt{d'}}\right) V_g, \tag{7}$$

where d' represents the feature dimension of key and query vectors. We further utilize a residual fusion approach to balance text prompts and customized guidance:

$$f_g' = f_s + \lambda f_g, \tag{8}$$

where λ is the fusion parameter. We can achieve fine-grained control over character features by encoding the reference image with the CtrlGAN Encoder. To further balance the text prompts and customized guidance, we inject customized guidance in the last $\beta \in [0, 1]$ part of the inference steps rather than in all steps.

3.3 ID Consistency across Frames

The above method successfully extracts ID information from the reference image and achieves fine-grained control. However, we observe that while each generated frame retains the character's identity, there is a lack of character consistency across frames. As shown in the left of 4, the character's clothing differs between frames. To ensure overall consistency in the generated story, we introduce the ID-Consis Attention. Specifically, since generating the current frame requires information from the previous frame to maintain coherence, we adopt a probabilistic latent mixing strategy inspired by previous work Ho & Salimans (2022); Zhou et al. (2024a). In our design, the latent representation from the previous frame is fused with that of the current frame at a certain probability, enabling ID consistency in a lightweight and efficient manner.

Formally, let $\mathcal{I} \in \mathbb{R}^{B\times N\times C}$ represent a batch of image features, where B, N, and C denote the batch size, the number of tokens per image, and the channel dimension, respectively. The self-attention function is defined as $\mathrm{Attention}(X_k, X_q, X_v)$, where X_k, X_q, and X_v are the key, query, and value matrices. In standard self-attention, each image feature I_i in $\mathcal{I}$ is independently processed as follows:

$$O_i = \mathrm{Attention}(Q_i, K_i, V_i), \tag{9}$$

where Q_i, K_i, and V_i are projections of I_i.

To establish interactions across images in a batch and maintain subject consistency, ICA mixes a subset of tokens S_i from other image features within the batch, where the proportion of tokens mixed is determined by the mixing probability ρ:

$$S_i = \mathcal{R}(\{I_j\}_{j\neq i, j\in[1,B]}, \rho), \tag{10}$$

where $\mathcal{R}$ denotes a random mixing function that takes as input the set of image features $\{I_j\}$ where the indices j are in the batch range $[1, B]$ but not equal to i, and the mixing probability ρ, and outputs the mixed token subset S_i. The mixed tokens S_i are then paired with the original image feature I_i to form a new token set P_i. Linear projections are applied to P_i to compute the updated key K_{Pi} and value V_{Pi} for ICA, while the original query Q_i remains unchanged. The self-attention operation is then computed as:

$$O_i = \mathrm{Attention}(Q_i, K_{Pi}, V_{Pi}). \tag{11}$$

Compared to existing methods focusing on self-attention computation, the innovation and effectiveness of ICA are analyzed in detail in the *Supplementary Material*.

4 Experiment

4.1 Implementation Details

Models and Datasets. We choose SDXL and SD v1.5 Rombach et al. (2022) as our base generator and implement BLIP2 Li et al. (2023) to generate captions for images. The training dataset comprises both real and synthetic data. Specifically, we use $70,000$ high-resolution facial images from FFHQ Karras (2019), $40,000$ full-body images from SHHQ Fu et al. (2022), and $70,000$ synthetic images generated by StyleGAN2 Karras et al. (2020) to enhance diversity.

Training Details. The model is trained on an NVIDIA A100 GPU with a batch size of 16. The training phase is divided into two stages. Firstly, we train the mapping network and the projection matrices W_q^g, W_k^g and W_v^g of all the newly integrated image cross-attention layers, aiming to align W_+ space vector with Stable Diffusion. Subsequently, to achieve better performance in more generalized scenarios, only the cross-attention layers are trained, the weight of mapping network keeps freezed in Stage 2. Please refer to *Supplementary Material* for more details.

Evaluation Details. We adopt representative facial images from CelebA Liu et al. (2015) and Mystyle Nitzan et al. (2022) as reference images. Following StoryDiffusion Zhou et al. (2024a), we utilize the proposed Story-aware Module (SaM) powered by GPT-4o Hurst et al. (2024) to randomly generate 40 story prompts for each reference image, resulting in a total of 3000 images for evaluation. We first review datasets from existing work on story visualization and observe that the dataset used in StorySalon Liu et al. (2024) is constructed by querying keywords related to storytelling for children. However, frames extracted from these videos are often blurry and lack clear reference subjects, making them unsuitable for the customized story visualization task. Therefore, we follow the dataset construction strategy of StoryDiffusion Zhou et al. (2024a) to build a dataset with thousands of entries. Specifically, we utilize the proposed Story-aware Module (SaM) powered by GPT-4o Hurst et al. (2024) to randomly generate 40 story prompts for each reference image, resulting in a total of $3,000$ images for evaluation. To address the deficiencies of existing datasets, we select images with clear and well-defined character subjects from CelebA Liu et al. (2015) and Mystyle Nitzan et al. (2022) as reference images. These images are specifically curated and tailored to meet the requirements of the customized story visualization task. By doing so, our constructed dataset achieves better alignment with the objectives of this task.

4.2 Qualitative Comparison

In this section, we conduct a qualitative analysis to demonstrate the high-quality customized story visualization capability of our method. We compare our approach with three representative methods that focus on ID preservation and story visualization, StoryDiffusion Zhou et al. (2024a), PhotoMaker Li et al. (2024b) and StoryMaker Zhou et al. (2024b). To validate our method's fine-grained control over character identity in alignment with textual descriptions, we focus on four representative attributes: age, smile, eye shape, and lip color. The results are shown in 3 and **??**. For more results on fine-grained control of character attributes, see the *Supplementary Material*.

As shown in fig. 3, all baseline methods struggle to achieve fine-grained control over character attributes. In terms of age, they produce characters with minimal visible changes, regardless of textual descriptions. Similarly, none of the baselines successfully reflects the gradual progression in smile intensity specified in the prompts (3, right). Regarding ID preservation and visual consistency, PhotoMaker maintains identity to some degree, but often generates incoherent images, such as inconsistent clothing across frames. StoryDiffusion exhibits poor ID preservation and relies heavily on detailed clothing prompts to maintain consistency. For example, the prompt ”casual sweater” leads to significant variation in outfit appearance. StoryMaker demonstrates excellent ID fidelity, yet tends to overfit to the reference image. This leads to a lack of variation in facial attributes, making fine-grained attribute control unachievable. Even worse, the generated images often fail to align with the textual description and instead merely replicate the reference face. For example, as shown in the yellow box (3, left), attributes such as eyeglasses are not correctly rendered according to the prompt.

In contrast, our proposed StoryCtrlnot only preserves identity and generates a coherent image sequence, but also enables fine-grained control over attributes in alignment with the text. As shown in **??**, attributes such as the degree of eye openness and lip color are accurately rendered according to the story description, demonstrating the effectiveness of our approach. We further conduct a **user study** to validate the effectiveness of our method (see detailed results in the *Supplementary Material*).

4.3 Quantitative Comparison

We present a comprehensive quantitative evaluation of different methods to evaluate the effectiveness of our method in customized story visualization. Specifically, we select four representative metrics: (1) ID Preservation: We use GroundingDINO Liu et al. (2025) with the input category ”face” to crop the facial regions from generated images. We then calculate the average CLIP Image Similarity Radford et al. (2021b) between the facial regions of each frame and the reference image. For a more comprehensive evaluation, we also use RetinaFace Deng et al. (2020) as the detection model and Arcface Deng et al. (2019) to extract the Face embedding. We then compute the face similarity by detecting and cropping the facial regions between the generated image and the reference image with the same ID. (2) Overall Consistency: We use GroundingDINO with the input category ”person” to crop character regions from generated images, then calculate the average pairwise CLIP Image Similarity between character regions across all images in a story. (3) Text Alignment: We calculate the CLIP Text Similarity Hessel et al. (2021) between each frame in the story and the prompt. As shown in 1, our method outperforms state-of-the-art methods across four metrics.

4.4 Ablation Study

Ablation on the CtrlGAN Encoder and SaM. To validate the effectiveness of our design, which combines the CtrlGAN Encoder with the Story-aware Module (SaM), in achieving fine-grained control over character attributes, we conduct an ablation study. We visualize the results in the left of **??**. Without the collaboration between the CtrlGAN Encoder and the Story-aware Module, the corresponding W_+ space vector w_r remains unchanged. As a result, the generated images fail to

Table 1: Quantitative comparison between our StoryCtrland other methods. **Bold** indicates the best result, and underline indicates the second best result.

Method	**Text Align.**↑ (%)	**Face Sim. (CLIP)** ↑ (%)	**Face Sim. (ArcFace)** ↑ (%)	**Overall Consis.** ↑ (%)
PhotoMaker-V2	29.8	81.5	68.2	68.9
StoryMaker	27.1	82.3	75.2	69.6
StoryDiffusion	29.9	78.1	64.5	74.3
Ours	**30.6**	**83.1**	**75.4**	**74.9**

Table 2: Quantitative ablation results about the mixing probability of our ID-Consis Attention.

	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$
Character Consistency ↑ (%)	75.1	75.8	76.4	77.5	75.5
Text Alignment ↑ (%)	33.7	33.9	33.8	33.8	33.8

reflect the attribute variations described in the text. In contrast, our full model, which integrates both components, successfully captures attribute changes described in the text, demonstrating fine-grained control.

Ablation on the ICA. To validate the effectiveness of the proposed ICA in improving visual consistency, we perform an ablation study. The right section of Fig.4 presents the results. As indicated by the yellow box, the absence of ICA leads to inconsistency in character clothing in the third frame, thereby compromising the overall visual coherence. Furthermore, we conduct quantitative experiments to validate our observations. The data presented in 3 indicate that lower ρ results in reduced ID consistency and overall image sequence coherence. In practice, we choose $\rho = 0.7$ to balance character consistency and text alignment.

We also perform qualitative analyses of the sampling rate ρ, with visual results provided in the *Supplementary Material*.

Customized Guidance Fusion. We conduct an ablation study on the fusion parameter λ to validate our ability to preserve ID identity. As shown in 7, when λ is small, the text prompt dominates, resulting in poor identity preservation of the generated character. Conversely, when λ is large, the lack of text prompt guidance prevents effective generation. We further conduct quantitative experiments. The results are shown in 3, which are consistent with our qualitative results. Our method sets λ to 0.8, achieving a balance between text prompt and customized guidance.

Injection Ratio. We conduct an ablation study on the injection ratio of customized guidance. Qualitative and quantitative results are in the *Supplementary Material*. We set $\beta = 0.8$ to balance text prompts and customized guidance.

5 Conclusion

In this paper, we present a novel framework for customized story visualization that focuses on fine-grained control aligned with narratives. Our method achieves fine-grained control over character attributes and high-fidelity ID preservation through the collaborative design of CtrlGAN and the Story-aware Module (SaM), while the ID-Consis Attention ensures the generation of coherent and consistent image sequences. Comprehensive qualitative and quantitative experiments, together with human preference evaluation, demonstrate that our proposed StoryCtrl outperforms current state-of-the-art methods in story visualization.

References

Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Table 3: Quantitative ablation results about the fusion parameter of our Customized Guidance Fusion.

	$\lambda = 0$	$\lambda = 0.4$	$\lambda = 0.8$	$\lambda = 2$
Face Similarity ↑ (%)	62.6	74.6	**76.6**	38.6

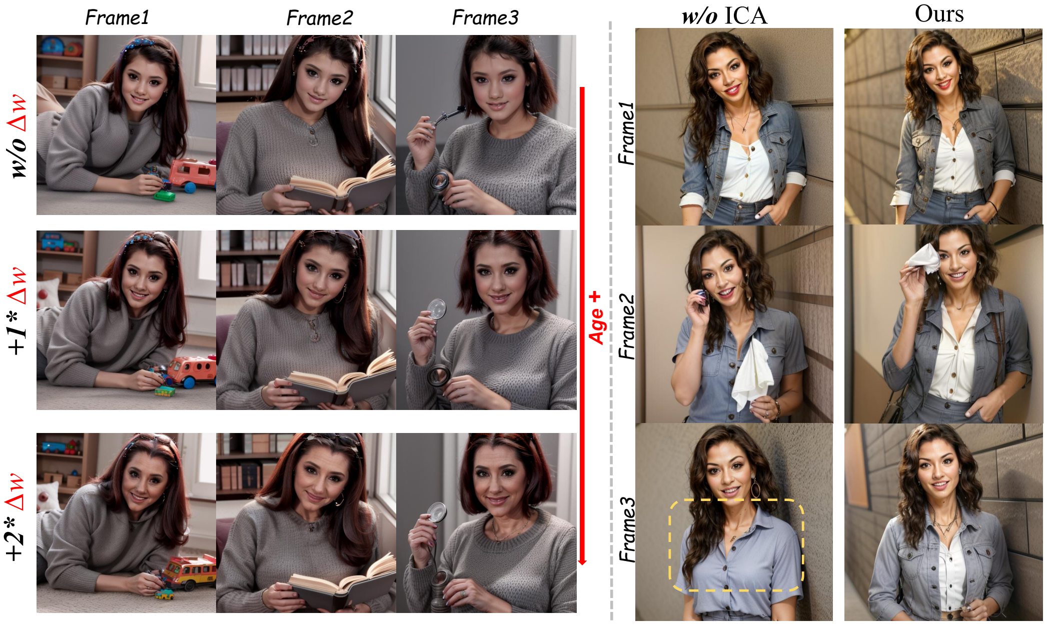


Figure 4: Qualitative ablation results about our CtrlGAN and SaM (left), ID-Consis Attention (right).

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6593–6602, 2024b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5203–5212, 2020.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, pp. 1–19. Springer, 2022.

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

Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, et al. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*, 2023.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*, 2023.

Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation, 2018. URL `https://arxiv.org/abs/1710.10196`.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Xiaoming Li, Xinyu Hou, and Chen Change Loy. When stylegan meets stable diffusion: a w+ adapter for personalized image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2187–2196, 2024a.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8640–8650, 2024b.

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

Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. A survey on leveraging pre-trained generative adversarial networks for image editing and restoration, 2022. URL `https://arxiv.org/abs/2207.10309`.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2025.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiawei Mao, Xiaoke Huang, Yunfei Xie, Yuanqi Chang, Mude Hui, Bingjie Xu, and Yuyin Zhou. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization, 2024.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)*, 41(6):1–10, 2022.

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, 2021a.

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, 2021b.

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*, pp. 10684–10695, 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9243–9252, 2020.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8543–8552, 2024.

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

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

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

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

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15943–15953, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, pp. 1–20, 2024.

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.

Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pp. 85–101. Springer, 2022.

Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *arXiv preprint arXiv:2411.17440*, 2024.

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*, 2024a.

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

Barret Zoph, Ekin D Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V Le. Learning data augmentation strategies for object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 566–583. Springer, 2020.

OVERVIEW

This supplementary material provides a detailed study of the proposed methodologies and their experimental validation. The main sections are summarized as follows:

- **Extended Discussion and Analysis of Proposed Methodologies**: An in-depth exploration of the theoretical framework and key design principles.
- **Dataset Construction and Comparison**: Describes the process of building datasets and compares their performance.
- **User Study**: Evaluates the usability and practical relevance of the proposed approach through user feedback.
- **Application**: Showcases potential applications of the proposed methodologies in real-world scenarios.
- **More Implementation Details**: Summarizes training protocols, hyperparameter settings, and computational requirements.
- **The Architecture of Mapping Network**: Provides detailed descriptions of the network architecture.
- **Additional Results**: Presents supplementary results to further validate the approach.

A EXTENDED DISCUSSION AND ANALYSIS OF PROPOSED METHODOLOGIES

A.1 DISCUSSION ON ID-CONSIS ATTENTION

Compared with related work Hong et al. (2023); Tewel et al. (2024); Zhou et al. (2024a) that focuses on self-attention mechanisms, the innovations of our proposed ID-Consis Attention (ICA) lie in the following aspects:

Probabilistic Reference Integration for Overall Consistency. Inspired by Classifier-Free Guidance (CFG), which significantly improves generation quality by interpolating between conditional and unconditional predictions without requiring additional classifier training, we draw a similar insight: Can temporal consistency be achieved without extra training or specific architectural modifications? Motivated by this, we propose an innovative design that probabilistically integrates reference information from preceding frames with independently generated frames that focus solely on the current context. This approach effectively balances overall consistency and independent frame generation while eliminating the need for additional training or custom designs. Furthermore, it ensures seamless compatibility with various models, enabling robust and coherent generation across frames without compromising flexibility or generality.

Flexible Frame Referencing. A distinguishing feature of our approach, compared to prior consistency-focused attention mechanisms, is the ability to flexibly control the number of referenced frames. This design effectively mitigates excessive memory consumption, addressing the common limitations of GPU resources in sequence generation tasks. By reducing computational overhead, our method enables scalable and efficient generation of extended sequences without sacrificing quality.

A.2 DISCUSSION ON CTRLGAN ENCODER

The motivation for introducing CtrlGAN into the diffusion space stems from addressing fine-grained control challenges in customized story generation tasks. While using detailed prompts can somewhat improve generation quality, this approach faces significant challenges in character customization tasks. As demonstrated by our *Qualitative Comparison* in the main paper, existing customization methods often result in generated images that appear overly fitted to the reference image, creating a visual effect akin to a pasted image. Merely using prompts proves insufficient for controlling complex generation processes.

Notably, StyleGAN's latent spaceKarras (2019) naturally decouples identity-irrelevant attributes, with the W_+ space capable of covering various fine-grained facial attributes, thereby addressing these challenges. Specifically, our CtrlGAN Encoder achieves feature disentanglement by dividing StyleGAN's W_+ space into its linear subspaces, each representing specific facial attributes. When a particular attribute in the story changes, the corresponding vector $\tilde{w}_r$ in the subspace is adjusted, with k tuned per story and Δw as a fixed hyperparameter.

B Dataset Construction and Comparison

We first review datasets from existing work on story visualization and observe that the dataset used in StorySalon Liu et al. (2024) is constructed by querying keywords related to storytelling for children. However, frames extracted from these videos are often blurry and lack clear reference subjects, making them unsuitable for the customized story visualization task. Therefore, we follow the dataset construction strategy of StoryDiffusion Zhou et al. (2024a) to build a dataset with thousands of entries. Specifically, we utilize the proposed Story-aware Module (SaM) powered by GPT-4o Hurst et al. (2024) to randomly generate 40 story prompts for each reference image, resulting in a total of $3,000$ images for evaluation. To address the deficiencies of existing datasets, we select images with clear and well-defined character subjects from CelebA Liu et al. (2015) and Mystyle Nitzan et al. (2022) as reference images. These images are specifically curated and tailored to meet the requirements of the customized story visualization task. By doing so, our constructed dataset achieves better alignment with the objectives of this task.

In the main paper, we compare our method with mainstream approaches widely used in the domains of story visualization and customized generation tasks to validate its effectiveness. Regarding additional related work, although the task of customized story visualization is distinct from both purely customization tasks and pure story visualization tasks, we further corroborate our results by comparing our method with additional customization-focused approaches. We reference StoryGen and use GPT-4o to generate more complex stories, resulting in a dataset with about 3k entries.

Method	Text Align.	Face Sim. (Clip)	Face Sim. (Arc.)	Consis.
StoryGen	24.6	59.2	46.6	39.8
ConsiStory	**31.6**	-	-	67.5
InstantID	25.9	76.8	67.9	<u>74.1</u>
IP-Adapter (FLUX)	28.5	<u>77.6</u>	<u>71.0</u>	68.9
Ours	<u>30.1</u>	**86.9**	**71.2**	**76.1**

Table 4: Performance Comparison of Various Methods on Text Alignment, Face Similarity, and Consistency. Best results are in **bold**; second-best are <u>underlined</u>.

Due to the limitations of the aforementioned dataset, StoryGen performs poorly on the customization task. ConsiStory lacks reference image support, and InstantID and IP-Adapter focus on faces, limiting text-aligned story generation. Overall, our method outperforms baselines and achieves competitive results.

C User Study

To validate the effectiveness of our method, we conduct a comprehensive user study. Due to the lack of reliable metrics for fine-grained attribute control, we design the study to evaluate three key dimensions: *fine-grained attribute control*, *identity preservation*, and *overall consistency*. Each of these dimensions is assessed through human subject evaluation to ensure the subjective quality of our method's outputs.

We recruit 14 participants with expertise in computer vision and generative AI for the study. The evaluation is based on 14 sets of results, covering 7 different character attributes. Each set includes outputs from our method alongside outputs from four state-of-the-art (SOTA) comparison methods to ensure fairness and reliability in the evaluation process. For each dimension, participants are asked to select the result they believe performs best.

As shown in Figure 5, our method achieves the highest user preference scores for fine-grained attribute control. Additionally, consistent with the results obtained from objective metrics in quantitative comparisons, our method also outperforms all baseline methods in terms of *identity preservation* and *overall consistency*. This demonstrates the robustness and effectiveness of our approach across multiple evaluation criteria.

D Application

As shown in Figure 6, our method also performs well in multi-character scenarios. Notably, despite the absence of specific multi-character settings in our training data, our approach demonstrates the

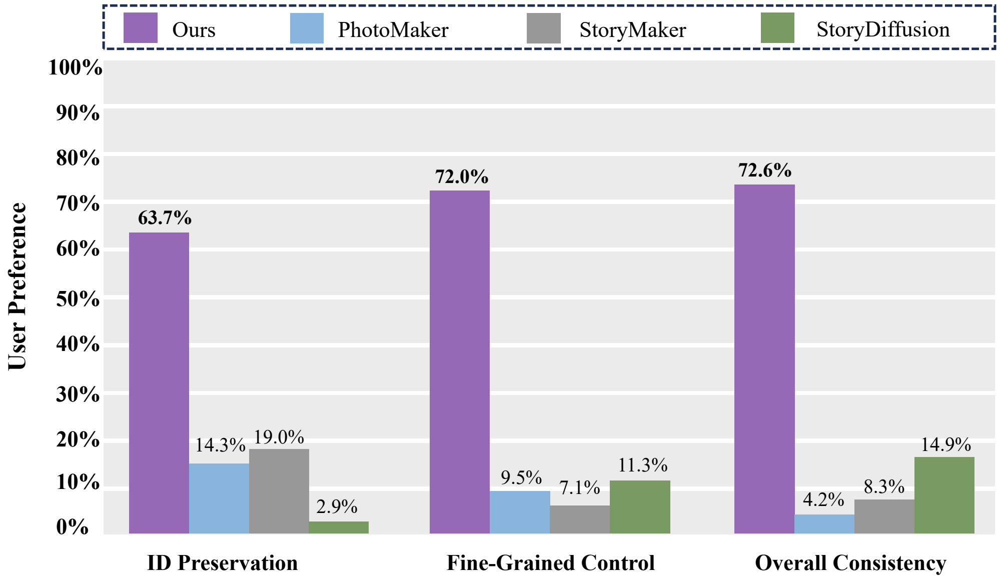


Figure 5: User study comparison between StoryCtrl and state-of-the-art (SOTA) methods on fine-grained attribute control, identity preservation, and overall consistency.

ability to achieve effective ID preservation across multiple characters. This is made possible through a simple inference modification, which involves concatenating ID embeddings from multiple inputs, extracting foreground masks for each individual using text cross-attention, and enhancing identity cross-attention with these masks. These modifications not only maintain identity consistency but also enable fine-grained control in scenarios involving multiple characters.

Furthermore, even in cases where the input text lacks explicit emotional descriptions, our method is capable of identifying subtle attribute trends within the narrative. This allows for fine-grained control over story visualization and ensures the generated content faithfully reflects nuanced changes in the underlying storyline.

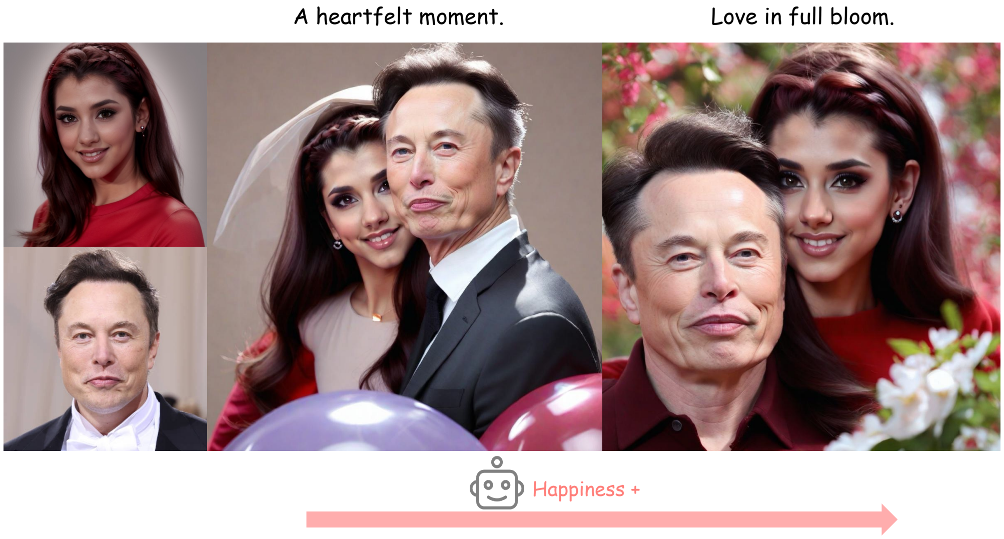


Figure 6: Visualization of our method's application in multi-character scenarios.

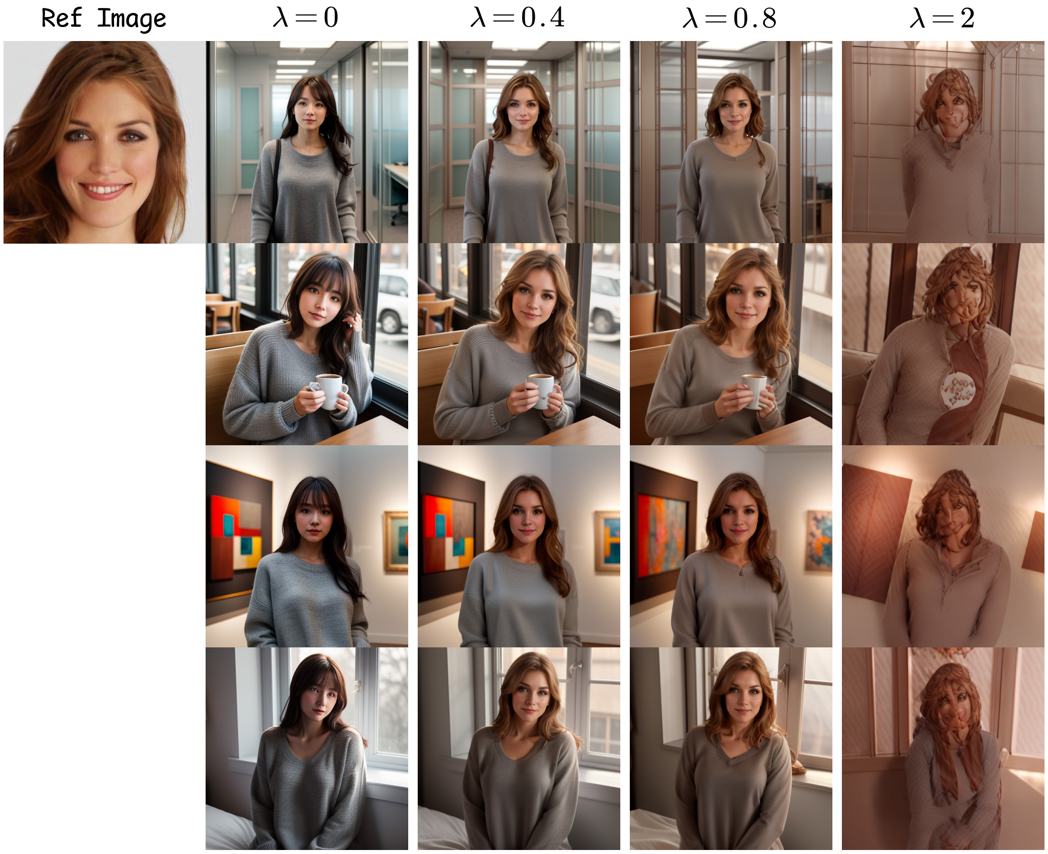


Figure 7: Qualitative ablation results about the fusion parameter of our Customized Guidance Fusion.

E Ablation Study

Injection Ratio. We conduct an ablation study to evaluate the effect of the injection ratio β for customized guidance. Qualitative and quantitative results are provided in the *Supplementary Material*. As shown in Figure 8, when β is high, the generated images tend to overfit the reference image (evidenced by artifacts, such as the face in the mural when $\beta = 1$). Conversely, when β is low, the ID fidelity decreases significantly. To balance the influence of text prompts and customized guidance, we select $\beta = 0.8$ as the optimal value for our final configuration.

Ablation on ICA. We further investigate the effect of the mixing probability ρ on ICA's performance. As shown in Table 5, lower ρ results in reduced ID consistency and less alignment of background details with the textual input. In practice, we select $\rho = 0.7$ as it offers the best trade-off between character consistency and text alignment, ensuring coherent and visually consistent outputs. Additionally, *Supplementary Material* includes qualitative results analyzing the influence of ρ.

Table 5: Quantitative ablation results on the mixing probability ρ for ICA.

Mixing Probability ρ	ID Consistency ↑	Text Alignment ↑
0.3	75.8	33.9
0.5	76.4	33.8
0.7 (Ours)	**77.5**	**33.8**
1.0	75.5	33.8

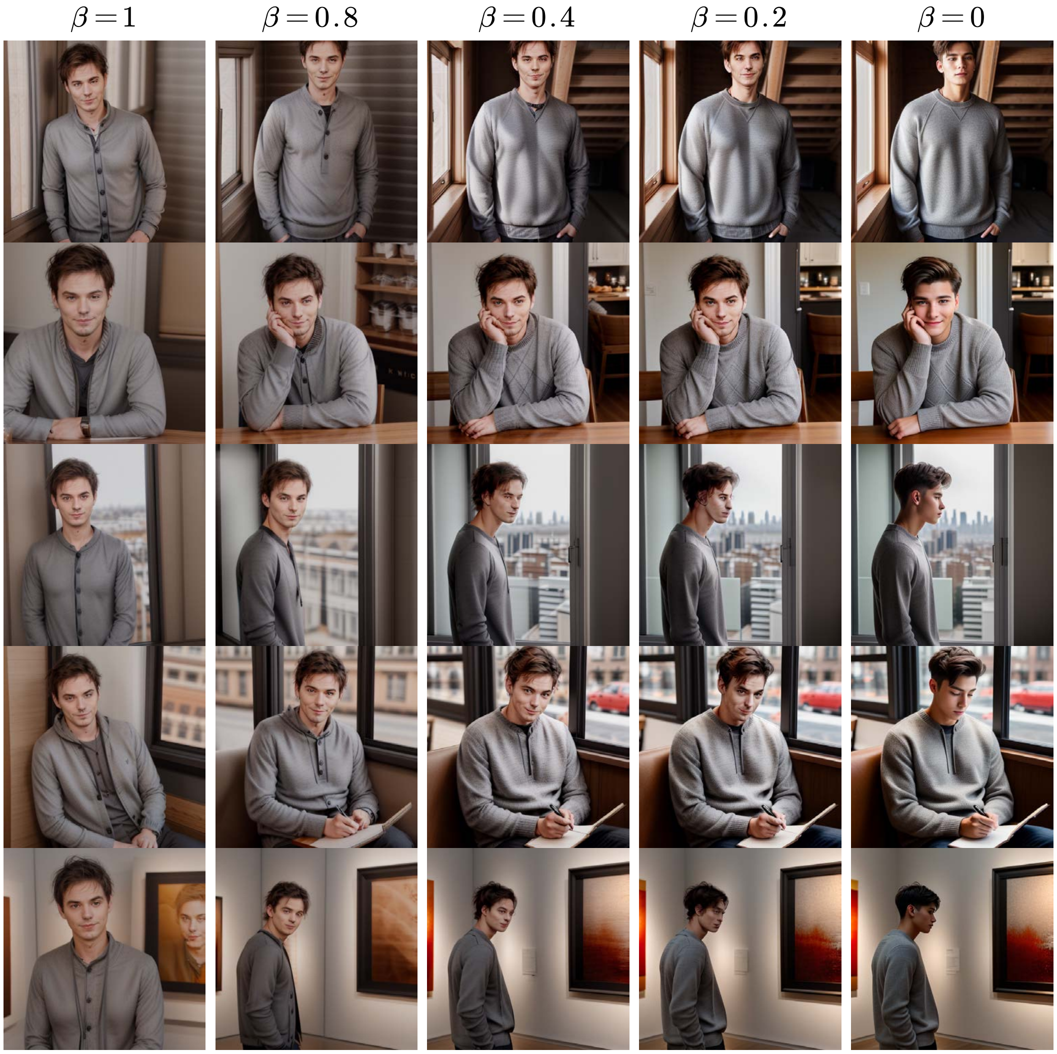


Figure 8: Qualitative results from the ablation study on the injection ratio β. Higher β leads to overfitting, while lower β causes ID fidelity degradation.

These findings demonstrate the importance of ICA and its configurations in fostering consistency across both individual frames and the overall sequence.

F More Implementation Details

We provide details of the training configuration used in our experiments for reproducibility:
Parameters. We set both λ and β to 0.8 to balance the influence of various components in our method.
Model Architecture. Following the training strategy of Li et al. (2024a), we utilize the standard e4e encoder Tov et al. (2021) as our CtrlGAN encoder. The mapping network consists of four Linear Layers with one LayerNorm. This design allows efficient and accurate feature mapping to better handle diverse input conditions.
Optimization. We adopt the AdamW optimizer Loshchilov & Hutter (2017) with a learning rate of 1×10^{-4} and a weight decay coefficient of 0.01. These hyperparameters are carefully chosen to ensure stable convergence during training.
Data Augmentation. To improve dataset diversity and variability, we implement several data augmentation techniques, including color jittering Zoph et al. (2020), stochastic rotation, and sampling of in-the-wild images. These augmentations help enhance the model's generalization ability and robustness in real-world scenarios.

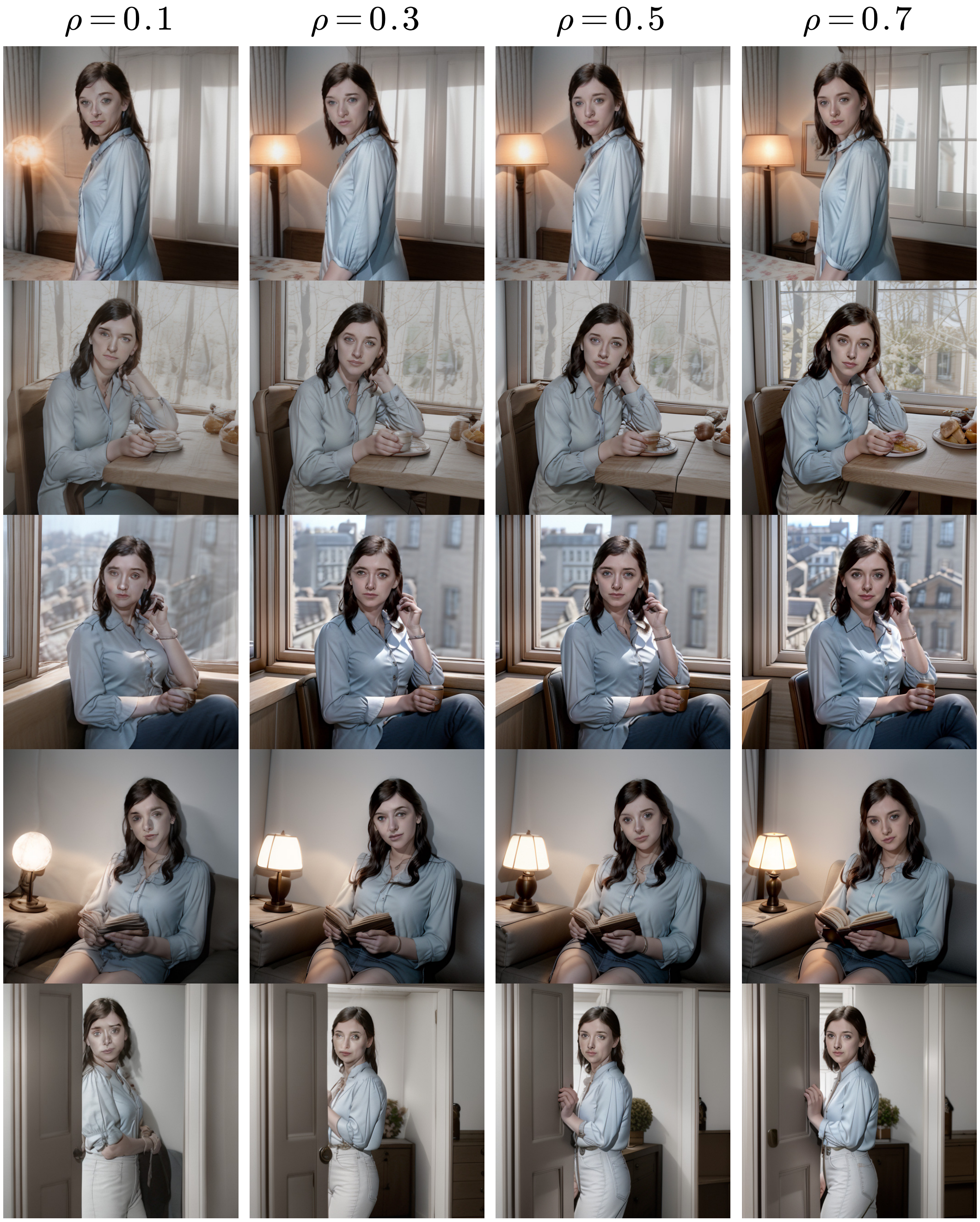


Figure 9: Qualitative ablation results about the mixing probability of ID-Consis Module.

Inference. During the inference phase, we employ a 50-step DDIM sampler Song et al. (2020). For classifier-free guidance Ho & Salimans (2022), we set the guidance scale to 7.5 to achieve a balance between fidelity and diversity in generated results. In addition, following IP-Adapter Ye et al. (2023), we apply a random dropout with a probability of 0.05 to both the text features and the W_+ space vector, which helps improve robustness and prevent overfitting.

These settings ensure effective training and inference for our method, allowing it to generalize well to diverse and challenging scenarios.

G The Architecture of Mapping Network

Due to the introduction of CtrlGAN as the image encoder, the resulting latent representation resides in the StyleGAN space with a dimension of 18×512. To leverage the generative capabilities of

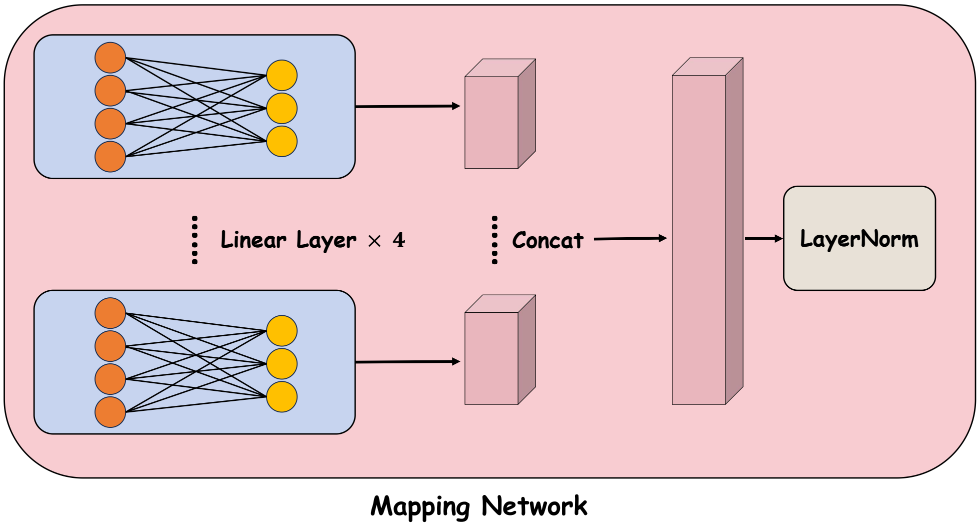


Figure 10: The architecture of Mapping Nerwork

the diffusion model, we employ a Mapping Network to project this latent into the Stable Diffusion, where the required input dimension is 4×768. As illustrated in Fig. 10, the image latent features in StleGAN space are passed through four linear layers, and the outputs are concatenated to form the final embedding, which is then fed into the diffusion model.

H Additional Results

In this section, additional qualitative comparisons are presented in Fig. 11 and Fig. 12, and a subset of them is further evaluated through a user study. Due to the lack of reliable metrics for fine-grained attribute control, we adopt user preference results to demonstrate the advantages of our method. The results show that our method outperforms existing state-of-the-art approaches in terms of perceived quality. To conduct a more comprehensive evaluation, we also assess identity preservation and overall consistency. These findings are consistent with our quantitative results, further confirming the effectiveness of our approach for customized story visualization tasks.

I Usage of Large Language Models

In this paper, large language models (GPT-4o) are used solely for polishing the writing of our manuscript. You may include other additional sections here.

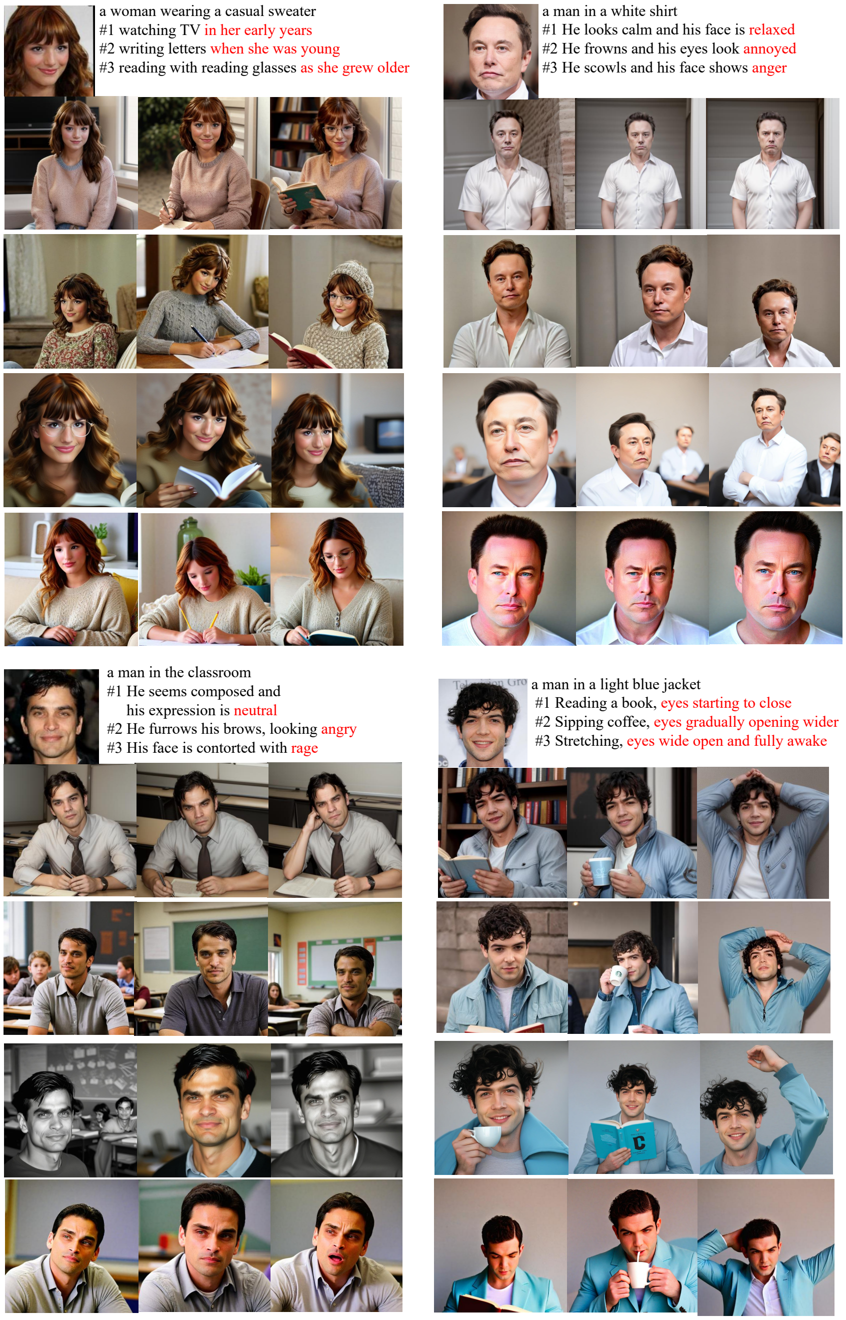

a woman wearing a casual sweater
#1 watching TV in her early years
#2 writing letters when she was young
#3 reading with reading glasses as she grew older
a man in a white shirt
#1 He looks calm and his face is relaxed
#2 He frowns and his eyes look annoyed
#3 He scowls and his face shows anger
a man in the classroom
#1 He seems composed and his expression is neutral
#2 He furrows his brows, looking angry
#3 His face is contorted with rage
a man in a light blue jacket
#1 Reading a book, eyes starting to close
#2 Sipping coffee, eyes gradually opening wider
#3 Stretching, eyes wide open and fully awake

Figure 11: Additional results about fine-grained control over attributes such as emotional expression (e.g., anger) and eye openness. From top to bottom: our StoryCtrl, PhotoMaker-V2, StoryMaker, and StoryDiffusion.

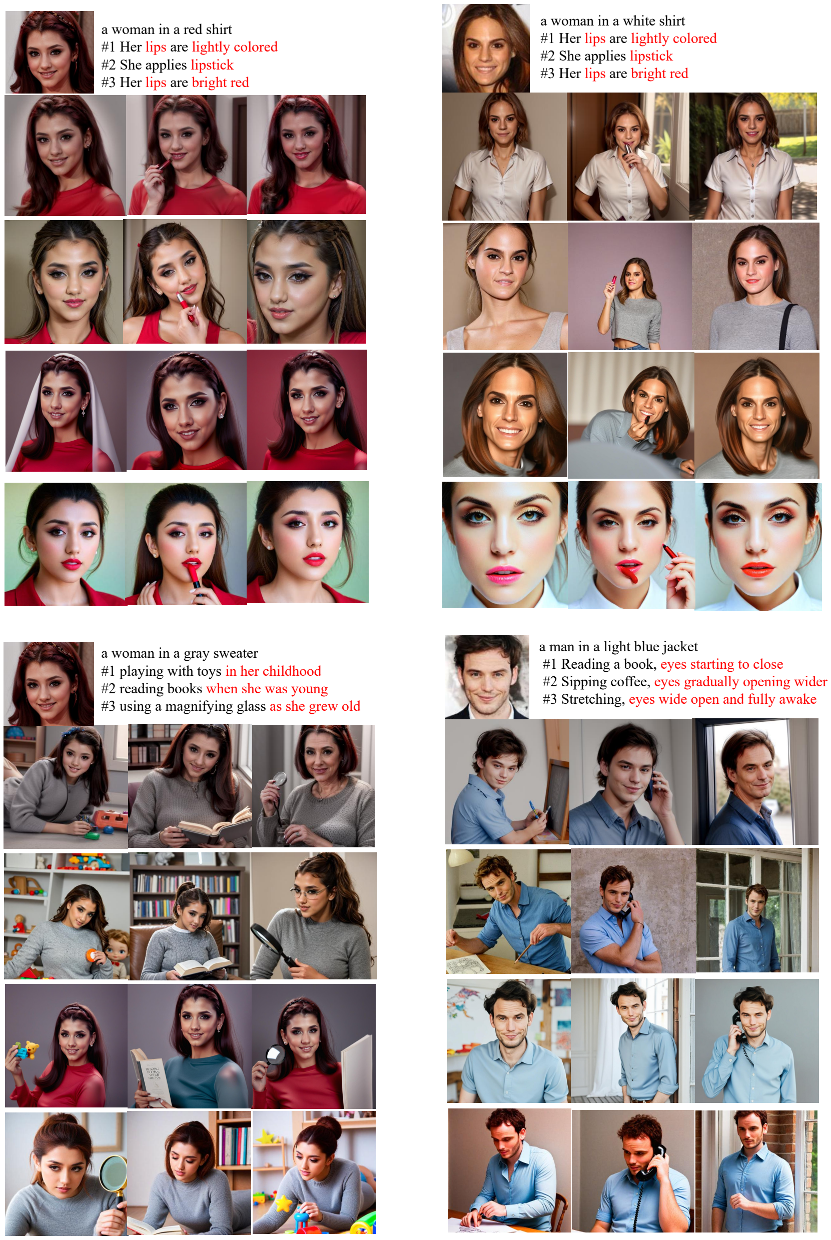

a woman in a red shirt
#1 Her lips are lightly colored
#2 She applies lipstick
#3 Her lips are bright red
a woman in a white shirt
#1 Her lips are lightly colored
#2 She applies lipstick
#3 Her lips are bright red
a woman in a gray sweater
#1 playing with toys in her childhood
#2 reading books when she was young
#3 using a magnifying glass as she grew old
a man in a light blue jacket
#1 Reading a book, eyes starting to close
#2 Sipping coffee, eyes gradually opening wider
#3 Stretching, eyes wide open and fully awake

Figure 12: Additional results about fine-grained control over attributes such as lip color and age. From top to bottom: our StoryCtrl, PhotoMaker-V2, StoryMaker, and StoryDiffusion.