

EvoTale: Continual Character Customization for Expanding Story Worlds

Jinlu Zhang¹, Qiyun Wang¹, Baoxiang Du¹, Jiayi Ji¹, Jing He²,
Rongsheng Zhang³, Tangjie Lv³, Xiaoshuai Sun^{1*}, Rongrong Ji¹

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, 361005, P.R., China.

²The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, 510000, P.R., China.

³Fuxi AI Lab, NetEase Inc., Hangzhou, 310000, P.R., China.

Abstract

Character-centric story visualization aims to synthesize coherent image sequences that depict narrative events and interactions while preserving recurring character identities. In expanding story worlds, new user-specified characters must be continually incorporated despite varying customization difficulty and identity conflicts in multi-character scenes, without disrupting previously learned identities. In this paper, we propose EvoTale, a continual character customization framework for expanding stylized story worlds. We first introduce an All-in-One-World Character Integrator, which accumulates character-specific residual components within a unified LoRA branch using sparsely overlapping subspaces spanned by a shared orthonormal basis, while freezing previously learned components to limit cross-character coupling. We then develop a Character Quality Gate that uses rubric-guided MLLM feedback as a bounded controller to adapt the optimization budget based on the assessed customization quality. Finally, we propose Character-Aware Region-Focus Sampling, which combines bounding-box-guided regional denoising with identity-aware global denoising to preserve character identities within their designated regions while maintaining global narrative coherence. Experimental results show that EvoTale achieves a favorable balance across character fidelity, continual identity retention, multi-character generation quality, and story-text alignment compared with representative story visualization and customization methods.

Keywords: Image Generation, Diffusion Model, Story Visualization, Continual Character Customization

1 Introduction

Story visualization is an important task in computer vision with broad application potential across creative media production, visual content authoring, and visual communication (Li et al. 2019; Song et al. 2020; Tong et al. 2018; Bugliarello et al. 2023; Huang et al. 2024; Ma et al. 2012). Given a multi-sentence narrative,

story visualization aims to synthesize a sequence of story frames that faithfully depict the events, scenes, and character interactions described in the text while preserving the visual identities of recurring characters across frames (Li et al. 2019; Zhou et al. 2024; Zhang et al. 2025). A key challenge

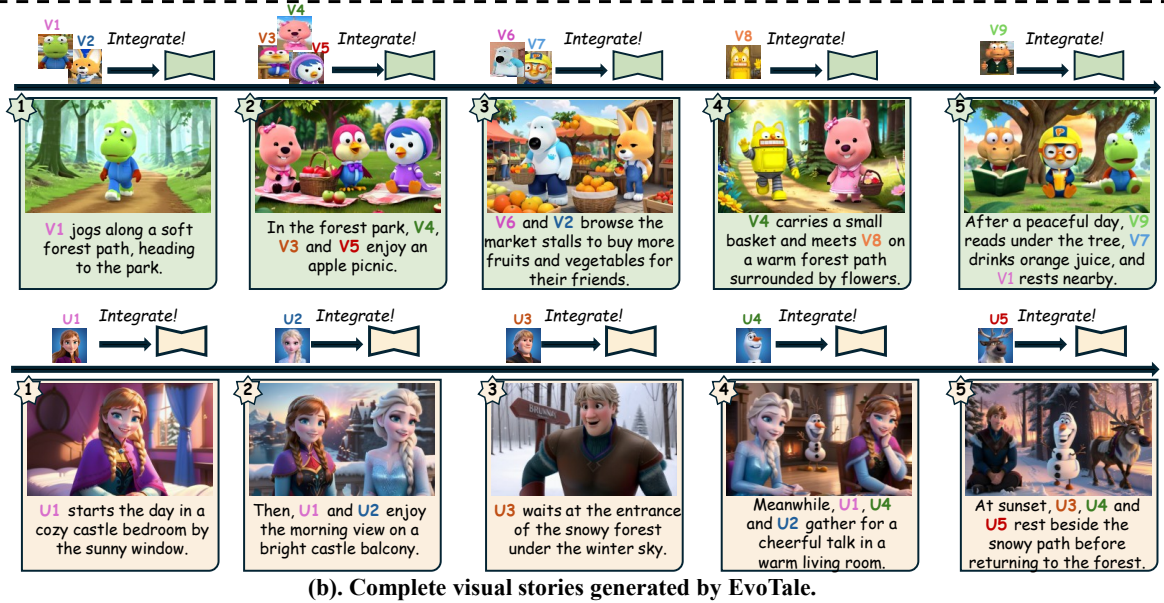
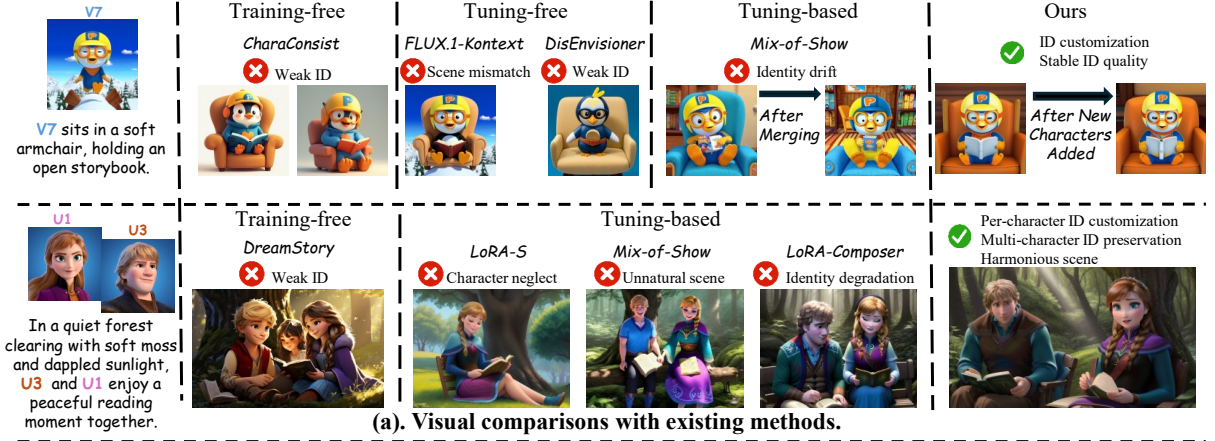


Fig. 1 Visual comparisons and continual character integration in expanding stylized story worlds. (a) Compared with representative methods, EvoTale better preserves user-specified identities and scene coherence in single- and multi-character settings. (b) EvoTale continually expands the character set while retaining previously integrated identities, as illustrated by V2 after adding V7 in *Pororo* and U2 after adding U4 in *Frozen*.

is therefore to jointly achieve accurate story-text alignment and cross-frame character consistency (Zhang et al. 2025; Dong et al. 2025; Zhuang et al. 2026; Pan et al. 2024), which becomes more demanding when new user-specified characters are incorporated sequentially into existing narratives, since each new character must be integrated without compromising previously learned character identities (Wang et al. 2024c; Dong et al. 2024; Guo and Jin 2025).

Existing story visualization methods can be broadly categorized into two types. Open-ended

methods (Liu et al. 2024; He et al. 2025a; Zhou et al. 2024; Jeong et al. 2023; Wang et al. 2025) generate visual narratives without a predefined cast, typically conditioning on narrative text, reference images, or internally generated cross-frame context. Though flexible across unseen stories and characters, they provide limited control over fine-grained user-specified characters. As a result, semantically plausible frames may deviate from the specified identities, as shown by the training-free results in Fig. 1(a). Closed-world methods (Li et al. 2019; Song et al. 2020; Chen et al. 2022;

Li 2022; Zhang et al. 2025; Tao et al. 2025; Shen and Elhoseiny 2025) are trained or adapted for a predefined cast, enabling stable character representations within a specific story world. However, extending the cast after training generally requires additional model adaptation.

A natural alternative is to incorporate image customization methods (Ruiz et al. 2023; Gal et al. 2023; Yang et al. 2025; Smith et al. 2023; Guo and Jin 2025; Dong et al. 2024; Labs et al. 2025) into story generation. However, existing approaches address only subsets of the requirements for expanding story worlds. Reference-conditioned methods (Labs et al. 2025; Ye et al. 2023; He et al. 2025b) inject identities without per-character optimization, but may lose fine-grained details or remain overly tied to reference appearances under changes in pose, viewpoint, scene, and interaction. Multi-concept customization methods (Gu et al. 2023; Kumari et al. 2023; Yang et al. 2025) support the composition of multiple customized concepts, but may suffer from identity drift, character neglect, or scene incoherence, as illustrated by the tuning-free and tuning-based results in Fig. 1(a). Continual customization methods (Dong et al. 2024; Guo and Jin 2025; Smith et al. 2023) enable sequential concept integration while mitigating forgetting. Nevertheless, existing methods remain limited in handling varying customization difficulty across characters and maintaining narrative consistency in multi-character generation. A uniform adaptation budget may underfit difficult identities while overfitting simpler ones. A practical framework should therefore combine continual character integration, quality-aware adaptation, and coherent multi-character composition with reliable identity assignment and global narrative coherence.

To address these limitations, we propose EvoTale, a continual story character customization framework for expanding stylized story worlds. EvoTale continually integrates new user-specified characters while preserving previously learned identities and supporting their coherent composition in multi-character story frames. Specifically, we first develop an **All-in-One-World Character Integrator**, which maintains a unified LoRA branch and incrementally incorporates character-specific residual components into sparsely overlapping subspaces spanned by a

shared orthonormal basis. Previously learned components remain frozen, limiting cross-character coupling to shared basis directions. This unified parameterization reduces interference and eliminates the need for post-hoc merging of independently trained adapters. To accommodate differences in character customization difficulty, we further introduce a **Character Quality Gate**, which uses rubric-guided multimodal large language model (MLLM) assessment solely as a bounded training controller to evaluate adaptation quality and adjust the training budget when necessary. Finally, for story frames involving multiple customized characters, we propose **Character-Aware Region-Focus Sampling**. It uses coarse bounding boxes to guide character-specific denoising within local regions, while global denoising captures the overall scene composition and interactions among characters. This sampling strategy improves character assignment and identity preservation while maintaining scene coherence and global narrative semantics (Fig. 1(b)).

We evaluate EvoTale under a continual character-learning protocol in the *Pororo* and *Frozen* story worlds in TBC-Bench (Zhang et al. 2025), covering both single- and multi-character story visualization, and compare it against representative story visualization and concept customization methods (Hu et al. 2021; Gu et al. 2023; Yang et al. 2025; Gal et al. 2023; Dong et al. 2024; Zhong et al. 2024; He et al. 2025b; Labs et al. 2025; Zhang et al. 2025; Po et al. 2024; Wang et al. 2025; He et al. 2025a). In SD1.5, EvoTale achieves the highest single-character D-I scores in both story worlds, with gains of 4.71% on *Pororo* and 3.77% on *Frozen*. It also exhibits limited identity degradation as characters are integrated sequentially. For multi-character generation, EvoTale achieves a favorable balance across character alignment, semantic alignment, and visual storytelling quality.

In summary, the contributions of this paper are threefold:

- We propose EvoTale, a framework for continual story character customization in expanding stylized story worlds, enabling the integration of new user-specified characters while preserving previously integrated identities.

- We develop an **All-in-One-World Character Integrator** with orthogonal-basis-guided residual accumulation and character-conditioned routing, a **Character Quality Gate** for quality-aware adaptation and **Character-Aware Region-Focus Sampling** for identity-preserving and globally coherent multi-character story generation.
- Extensive experiments show that EvoTale achieves a favorable balance across character fidelity, story-text alignment and continual identity retention.

2 Related Work

2.1 Story Visualization

Story visualization aims to synthesize coherent image sequences from multi-sentence narratives (Li et al. 2019; Tao et al. 2025; Shen and Elhoseiny 2025; Zhang et al. 2025; Zheng and Fu 2025). Existing methods can be broadly grouped into closed-world and open-ended paradigms according to whether they assume a predefined character set. Closed-world methods (Li et al. 2019; Tao et al. 2025; Shen and Elhoseiny 2025; Zhang et al. 2025; Zheng and Fu 2025) are trained or adapted for a particular story world and do not natively support persistent expansion to arbitrary user-specified characters, and extending the cast generally requires additional adaptation or retraining. Open-ended methods (Liu et al. 2024; He et al. 2025a; Zhou et al. 2024, 2025) condition on narrative text, reference images, or cross-frame context to support generation across unseen characters. However, these methods mainly preserve character consistency within the current story generation process, without persistently integrating newly introduced character identities for reuse in future stories. In contrast, EvoTale targets continual character customization for stylized story visualization, allowing the cast to expand while retaining previously integrated identities.

2.2 Image Customization

Image customization aims to adapt pretrained text-to-image models to faithfully synthesize user-specified visual concepts from only a few reference images (Ruiz et al. 2023; Gal et al. 2023). Tuning-free methods (Wei et al. 2023; He et al.

2025b; Wang et al. 2024b; Xiao et al. 2025; Ye et al. 2023; Labs et al. 2025) condition generation on reference images without subject-specific test-time optimization. Although efficient, they may either lose fine-grained identity details or inherit reference-specific appearance cues, limiting flexibility under substantial changes in pose, viewpoint, scene, or interaction. Tuning-based methods instead optimize subject- or domain-specific representations or parameters from limited reference data (Ruiz et al. 2023; Gal et al. 2023; Zhu et al. 2025), allowing fine-grained visual characteristics to be encoded through optimization. While effective for individual concepts, extending tuning-based customization to multiple concepts often requires joint or repeated optimization, or additional model-composition mechanisms (Gu et al. 2023; Kumari et al. 2023).

Multi-Concept Customization.

Multi-concept customization extends personalized image generation to jointly synthesize multiple user-specified concepts. Existing methods rely on joint optimization (Kumari et al. 2023; Han et al. 2023), post-hoc model merging (Gu et al. 2023; Po et al. 2024; Hoang et al. 2025), or inference-time composition (Kwon et al. 2024; Kong et al. 2025; Yang et al. 2025). Orthogonal Adaptation (Po et al. 2024), for example, encourages independently learned weight residuals to be mutually orthogonal, reducing interference during model aggregation. However, such approaches still rely on independently trained concept-specific components and require an explicit aggregation or composition step. When customized concepts arrive sequentially, continual learning (CL) provides a natural paradigm for learning new knowledge while preserving previously learned capabilities (Wang et al. 2024a; Cha et al. 2021; Wang et al. 2022; Kong et al. 2023; Cotogni et al. 2025). Recent continual customization methods extend the CL paradigm to support sequential concept learning while mitigating interference with previously acquired concepts (Dong et al. 2024; Guo and Jin 2025; Smith et al. 2023). However, existing methods mainly focus on retaining previously learned concepts during continual customization, rather than adapting the learning process to the needs of each newly introduced concept. In contrast, EvoTale first

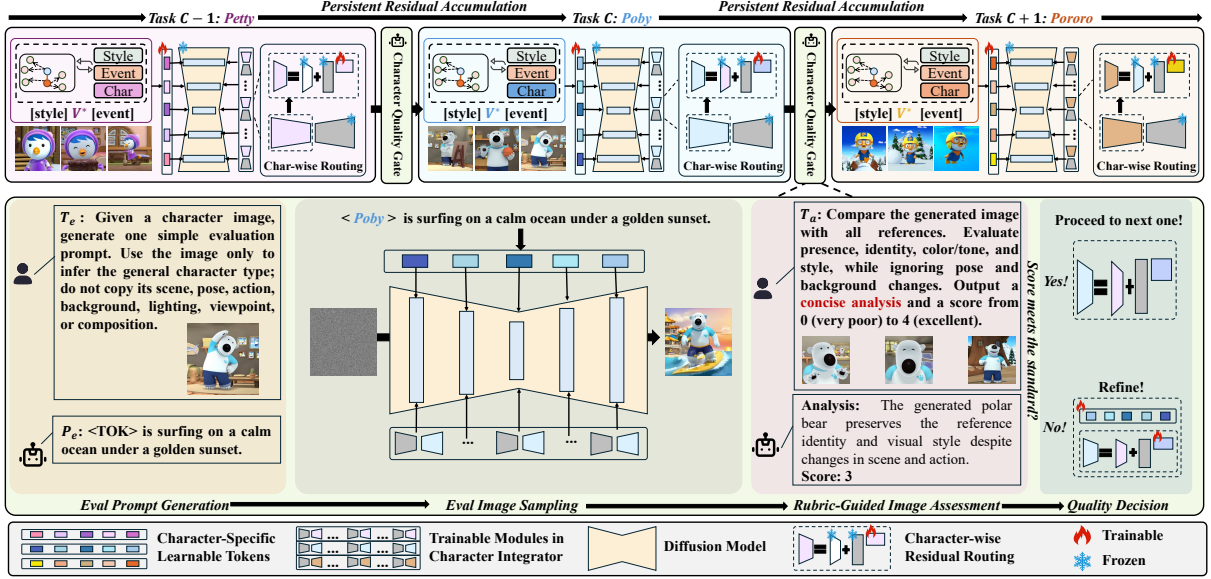


Fig. 2 Overview of EvoTale’s continual character customization pipeline. The All-in-One-World Character Integrator sequentially incorporates newly introduced characters into a unified framework while preserving previously acquired character knowledge. After each customization attempt, the Character Quality Gate (CQG) employs an MLLM as a quality judge to generate an evaluation prompt, assess the customized result, and determine whether to proceed to the next character or further refine the current one.

integrates new characters within a single persistent LoRA branch. The character-specific residual components are assigned to sparsely overlapping subspaces spanned by a shared orthonormal basis, while previously learned components remain frozen, limiting cross-character coupling to shared basis directions. EvoTale further adjusts the customization budget according to the assessed quality of each character.

Beyond maintaining multiple customized concepts in a unified model, another challenge is to compose multiple customized characters within a single image, where concept neglect, identity mixing, and cross-concept interference often arise (Xiao et al. 2025; Gu et al. 2023; Yang et al. 2025; Jiang et al. 2025). Existing methods address these issues through inference-time composition or optimization (Zhong et al. 2024; Jin et al. 2025; Jiang et al. 2025), attention-based identity control (He et al. 2025a; Zhao et al. 2026; Banerjee et al. 2025), and explicit spatial guidance such as masks, key-poses, or bounding boxes (Zhang et al. 2025; Dong et al. 2024; Gu et al. 2023; Banerjee et al. 2025). These mechanisms are often combined to improve concept separation and identity assignment. Story visualization further requires

multi-character composition to reflect narrative context and inter-character relations (He et al. 2025a; Zhang et al. 2025; Zhou et al. 2025; Banerjee et al. 2025). EvoTale uses coarse bounding boxes to guide character-specific regional denoising, while global denoising captures the overall scene structure and inter-character interactions. This design improves identity assignment and preservation while maintaining scene coherence.

3 Method

3.1 Overview

EvoTale adopts the Character-Graph from StoryWeaver (Zhang et al. 2025), with $G = \langle O, E, A \rangle$, where O , E , and A denote characters, events, and associated attributes. We retain this semantic representation and focus on continual character customization for an expanding character cast. Unlike StoryWeaver, whose generator is optimized for a predefined character set, EvoTale sequentially acquires new character identities while preserving previously learned ones.

As illustrated in Fig. 2 and Fig. 3, EvoTale consists of three complementary components.

First, the All-in-One-World Character Integrator (**CI**) accumulates character-specific residuals in a persistent LoRA branch and uses character-conditioned subspace routing during customization and inference. Second, the Character Quality Gate (**CQG**) uses an MLLM-based quality signal to accept the current customization or trigger bounded adaptive refinement. Finally, Character-Aware Region-Focus (**CA-RF**) Sampling combines Role-Specific Crop Denoising (**RS-CD**) for identity anchoring with Identity-Aware Global Denoising (**IA-GD**) for global harmonization through time-aware fusion.

3.2 All-in-One-World Character Integrator

Following prior methods (Gu et al. 2023; Po et al. 2024; Dong et al. 2024), we adopt LoRA (Hu et al. 2021) to encode character-specific visual knowledge. For a linear layer with weight W , LoRA introduces a low-rank residual $\Delta W = AB$, where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ denote the down- and up-projection matrices, respectively. Unlike conventional subject customization, continual story character customization must incorporate new characters as the cast expands while preserving previously integrated identities for reuse in subsequent story frames. This motivates a residual parameterization that supports continual accumulation with limited cross-character interference.

Existing multi-concept methods typically learn separate concept-specific modules and combine them when needed (Po et al. 2024; Hoang et al. 2025; Gu et al. 2023), whereas continual customization methods sequentially incorporate new concepts while preserving previously learned ones (Dong et al. 2024). In particular, Orthogonal Adaptation (Po et al. 2024) trains separate concept-specific residuals using bases sampled from a shared orthogonal matrix and aggregates the resulting modules after customization. Building on its randomized orthogonal-basis principle, our All-in-One-World Character Integrator (**CI**) instead organizes all character-specific increments within a unified LoRA branch.

Specifically, for each LoRA-injected linear layer, we construct a fixed orthonormal basis $Q \in \mathbb{R}^{d \times d}$ by applying QR decomposition to a matrix whose entries are independently sampled from a standard Gaussian distribution, such that

$Q^\top Q = I_d$. For each incoming character c_i , CI randomly and uniformly samples r_1 distinct column indices to form an index set $\mathcal{S}_i$ and defines the corresponding character-specific basis as:

$$D_i = Q[:, \mathcal{S}_i], \quad P_i = D_i D_i^\top, \quad |\mathcal{S}_i| = r_1, \quad (1)$$

where $D_i^\top D_i = I_{r_1}$, and P_i denotes the projection matrix associated with c_i . The shared basis Q remains fixed throughout continual customization, while each D_i selects a character-specific subset of its basis directions. For two independently sampled subsets $\mathcal{S}_i$ and $\mathcal{S}_j$, the expected overlap is $\mathbb{E}[|\mathcal{S}_i \cap \mathcal{S}_j|] = r_1^2/d$, corresponding to a fraction of r_1/d relative to each character subspace. Therefore, when $r_1 \ll d$, the character subspaces share only a small fraction of their assigned directions.

CI accumulates character-specific knowledge in the down-projection factor A , while the shared up-projection factor B remains fixed. After integrating i characters, the accumulated factor is:

$$A^{(i)} = \sum_{k=1}^i D_k L_k, \quad \Delta W^{(i)} = A^{(i)} B, \quad (2)$$

where $L_k \in \mathbb{R}^{r_1 \times r}$ denotes the learnable coefficient matrix for character c_k . Each character contributes an additive component $D_k L_k$ to the unified LoRA branch. When a new character c_i is incorporated, its component $D_i L_i$ is added to $A^{(i-1)}$, while all previously learned components remain fixed. The shared matrix B is initialized once with entries sampled from a standard Gaussian distribution and remains fixed throughout continual customization.

Following (Gu et al. 2023; Dong et al. 2024), each character c_j is also associated with a set of layer-wise learnable character embeddings $\{v_j^\ell\}_{\ell=1}^L$. When customizing c_j , L_j and its character embeddings are jointly optimized, while the shared matrix B , all previous coefficients $\{L_k\}_{k < j}$, and previously learned character embeddings remain fixed. CI then routes the accumulated LoRA residual through the character-specific projector P_j , yielding:

$$XW + X P_j \left(A^{(j-1)} + D_j L_j \right) B. \quad (3)$$

Using $P_j = D_j D_j^\top$, $D_j^\top D_j = I_{r_1}$, and $A^{(j-1)} = \sum_{k < j} D_k L_k$, the routed residual is formulated as:

$$\begin{aligned} & X P_j \left(A^{(j-1)} + D_j L_j \right) B \\ &= X D_j \left(D_j^\top A^{(j-1)} + L_j \right) B \\ &= X D_j \left(L_j + \sum_{k < j} D_j^\top D_k L_k \right) B. \end{aligned} \quad (4)$$

In Equ.4, the first term captures the character-specific residual learned for c_j , while the second term represents residual contributions from previously integrated characters that are routed into the subspace of c_j . Since the shared matrix B and all previous coefficients $\{L_k\}_{k < j}$ remain fixed, optimizing L_j cannot directly overwrite their stored residual components. Moreover, $D_j^\top D_k$ has nonzero entries only for basis directions shared by the two characters. Contributions from previous characters are therefore restricted to the sparse overlap between their assigned subspaces, allowing the new character to be learned primarily through its own basis directions.

After customizing c_j , its learned component is added to the unified LoRA branch:

$$A^{(j)} = A^{(j-1)} + D_j L_j = \sum_{k=1}^j D_k L_k. \quad (5)$$

Both L_j and its character embeddings are then fixed during all subsequent customization steps.

We next examine the reverse direction, namely the influence of a newly integrated character on previously learned characters. After incorporating c_j , the residual routed for an earlier character c_i , where $i < j$, is formulated as:

$$\begin{aligned} & X P_i A^{(j)} B \\ &= X P_i A^{(j-1)} B + X P_i D_j L_j B \\ &= X P_i A^{(j-1)} B + X D_i \left(D_i^\top D_j \right) L_j B. \end{aligned} \quad (6)$$

The first term preserves the residual routed to c_i before learning c_j , whereas the second represents the additional influence introduced by the new character. This additional term is restricted by $D_i^\top D_j$ to the basis directions jointly selected by the two characters. CI therefore combines parameter freezing with sparse, basis-dependent coupling

to limit bidirectional cross-character interference, consistent with the limited identity degradation, as shown in Sec. 4.3.

After integrating T characters, generating a target character c_i uses the same character-specific subspace to route the unified LoRA residual:

$$XW + (X D_i) \left(D_i^\top A^{(T)} \right) B. \quad (7)$$

This factorized form is equivalent to $XW + X P_i A^{(T)} B$, but avoids explicitly constructing the $d \times d$ projection matrix P_i . The routing is confined to the LoRA residual branch and does not modify the pretrained weight W .

In our implementation, CI is applied to the query, key, and value projections of self-attention layers and to the query projection of cross-attention layers. For clarity, we omit layer indices throughout the formulation.

3.3 Character Quality Gate via MLLM-as-Judge

Another challenge in continual character customization is determining whether the current character has been sufficiently customized before moving on. A fixed training schedule provides no explicit quality signal for this decision. Moreover, characters with different shapes, accessories, and visual styles may require different amounts of optimization. Consequently, a uniform training configuration may under-customize difficult identities while over-optimizing simpler ones. To address this issue, we introduce a Character Quality Gate (**CQG**), a bounded quality-aware controller that evaluates each customization attempt and determines whether to accept the current character or trigger further refinement with adjusted learning rates and training steps.

At the end of each customization attempt for character c_i , EvoTale randomly samples n reference images $\mathcal{I}_i = \{I_i^j\}_{j=1}^n$ from the training images of c_i in each of the evaluation trials. One image from $\mathcal{I}_i$, together with a prompt-generation instruction T_e , is provided to an MLLM (Bai et al. 2025) to generate an evaluation prompt P_e for subsequent image generation. The instruction asks the MLLM to construct a prompt that places the character in a new scene, action, or environment. The customized model then generates an evaluation image I_e conditioned on P_e .

Inspired by DreamBench++ (Peng et al. 2025), we also use the MLLM as a quality judge for character customization. The MLLM compares I_e with all images in $\mathcal{I}_i$ under a rubric-guided instruction T_a and assigns an integer score $s \in [0, 4]$, corresponding to five quality levels from *very poor* to *excellent*. Our character-focused rubric considers four aspects, *i.e.*, *character presence*, *character identity*, *color and tone consistency*, and *style representation*. The score serves only as an internal control signal for CQG and is neither treated as an absolute measure of customization quality nor used as a final evaluation metric.

To stabilize this quality signal, CQG repeats the complete evaluation procedure K times and computes the mean score for the m -th customization attempt as $\bar{s}^{(m)} = \frac{1}{K} \sum_{\ell=1}^K s_{\ell}^{(m)}$. The mean score is compared against a predefined threshold τ . If $\bar{s}^{(m)} \geq \tau$, the current customization is accepted, and training proceeds to the next character. Otherwise, CQG triggers another refinement attempt and computes a score-adaptive refinement factor:

$$\rho^{(m)} = \max\left(\rho_{\min}, 1 - \exp\left(\bar{s}^{(m)} - \tau\right)\right). \quad (8)$$

The learning rate and number of training steps for the next attempt are updated as:

$$\begin{cases} lr^{(m+1)} = \max\left(lr_{\min}, \rho^{(m)} lr^{(m)}\right), \\ step^{(m+1)} = \max\left(step_{\min}, \rho^{(m)} step^{(m)}\right). \end{cases} \quad (9)$$

Here, $lr_{\min}$ and $step_{\min}$ are derived from the initial customization configuration using the same lower-bound factor $\rho_{\min}$, rather than introduced as additional tuning parameters. Since the refinement factor is applied recursively across successive attempts, these bounds prevent the optimization budget from vanishing after repeated scaling. A lower quality score produces a larger $\rho^{(m)}$, retaining more of the current optimization scale for further customization, whereas a score closer to τ results in a lighter refinement. This yields a bounded coarse-to-fine schedule in which refinement becomes progressively more conservative without degenerating into negligible updates.

Finally, CQG allows at most M customization attempts. The threshold τ therefore serves as an early-acceptance target rather than a quality

guarantee. The current customization is accepted once $\bar{s}^{(m)} \geq \tau$. Otherwise, if the threshold is not reached within M attempts, CQG retains the latest state and proceeds to the next character. This bounded design limits training overhead while allowing underperforming characters to receive controlled refinement.

3.4 Character-Aware Region-Focus Sampling

Generating multiple customized characters in a single story frame requires simultaneously preserving individual identities and maintaining a coherent global composition. A straightforward strategy is to combine role-specific crop predictions with standard full-image denoising using a fixed fusion weight. Although this provides strong local identity guidance, it may lead to rigid character layouts. Replacing the fixed weight with a time-aware schedule improves global composition, but the increasing contribution of standard global denoising may weaken identity-specific guidance because it lacks character-aware routing. These two strategies are illustrated by the first two settings in Fig. 8(b).

To balance these objectives, we introduce Character-Aware Region-Focus (**CA-RF**) Sampling, which performs two complementary denoising processes in parallel from the same noisy latent as shown in Fig. 3. A *Role-Specific Crop Denoising* (**RS-CD**) branch provides strong local identity cues through single-character crop denoising, while an *Identity-Aware Global Denoising* (**IA-GD**) branch performs character-aware full-image denoising through identity-conditioned attention routing. Their predictions are fused with a time-aware schedule, progressively shifting from RS-CD-dominant identity anchoring to IA-GD-dominant global harmonization.

For a frame containing m characters, CA-RF takes $\{\mathcal{P}_g, \{\mathcal{P}_i\}_{i=1}^m, \{\mathcal{B}_i\}_{i=1}^m\}$ as input, where $\mathcal{P}_g$ denotes the global story prompt, $\mathcal{P}_i$ is the character-specific prompt for c_i , and $\mathcal{B}_i = (h_i^s, w_i^s, h_i^e, w_i^e)$ denotes its bounding box. The global prompt provides the overall story context, while each character-specific prompt provides the identity condition for its corresponding character. The bounding boxes serve as coarse spatial priors that associate these character-specific conditions with their designated regions.

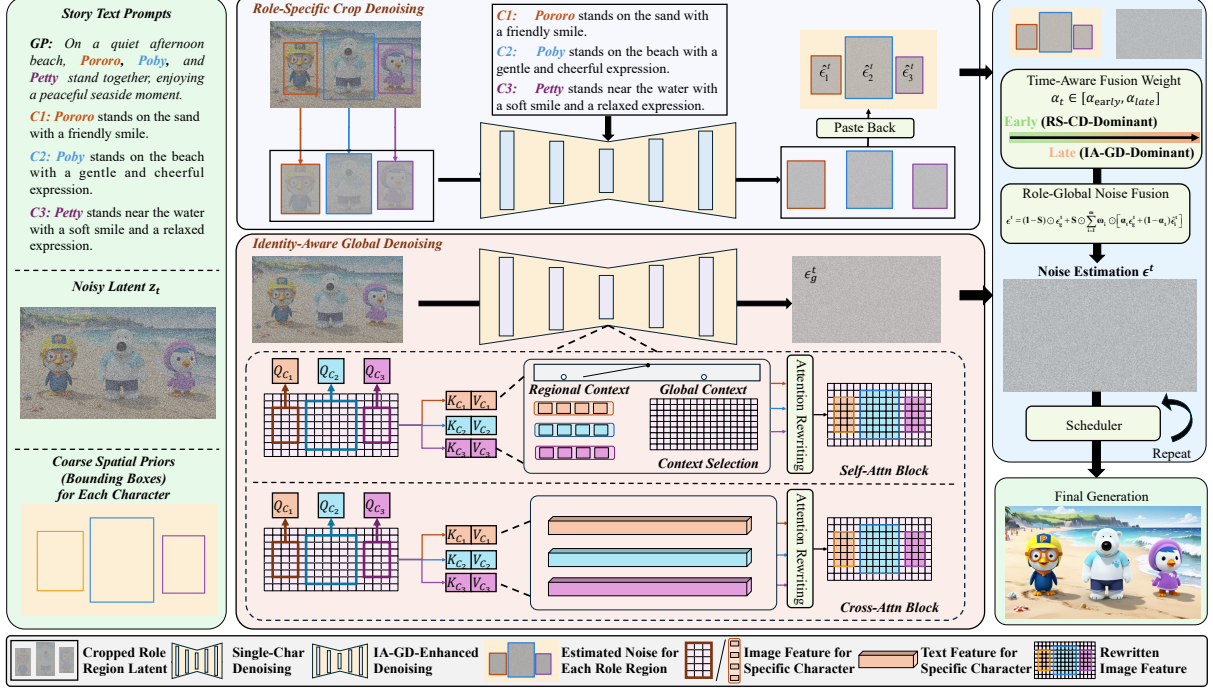


Fig. 3 Overview of Character-Aware Region-Focus Sampling. RS-CD predicts character-specific regional noise, while IA-GD preserves global context through character-conditioned attention. A time-aware schedule initially favors RS-CD for identity preservation and gradually shifts toward IA-GD for coherent scene composition.

Role-Specific Crop Denoising (RS-CD).

Inspired by the region noise estimation strategy in CIDM (Dong et al. 2024), RS-CD adopts a different crop-wise formulation that explicitly isolates each character during noise prediction. Rather than estimating region-conditioned noise on the full latent, RS-CD crops each character region from the current noisy latent and performs an independent single-character denoising forward.

For any tensor U with spatial resolution $H_U \times W_U$, we rescale the normalized bounding box $\mathcal{B}_i$ to the resolution of U and denote the result by:

$$\mathcal{B}_i^U = \text{Scale}(\mathcal{B}_i, H_U, W_U). \quad (10)$$

This notation applies to both latent-level cropping and feature-level regional processing.

Given the noisy latent z_t , RS-CD extracts the latent region associated with character c_i as:

$$z_{t,i} = \text{Crop}(z_t, \mathcal{B}_i^{z_t}), \quad (11)$$

and independently predicts the role-specific noise using the corresponding single-character condition

in CI routing:

$$\epsilon_i^t = \epsilon_\theta(z_{t,i}, t, \mathcal{P}_i, c_i). \quad (12)$$

The predicted crop noise is then mapped back to its original location on a full-size noise canvas, with zeros elsewhere:

$$\hat{\epsilon}_i^t = \text{Paste}(\epsilon_i^t, \mathcal{B}_i^{z_t}). \quad (13)$$

This procedure constructs a role-specific noise prediction, providing a strong character-specific signal for identity anchoring.

Identity-Aware Global Denoising (IA-GD).

To complement the isolated local predictions of RS-CD with full-image context, IA-GD first computes global attention and then enhances character regions using identity-conditioned attention. Let X denote an intermediate image feature. The standard attention mechanism is formulated as:

$$\text{Attn}(Q, K, V) = \sigma \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (14)$$

In this case, the global self- and cross-attention outputs are formulated as:

$$\begin{cases} O_g^s = \text{Attn}(XW_q^s, XW_k^s, XW_v^s), \\ O_g^c = \text{Attn}(XW_q^c, f_g W_k^c, f_g W_v^c), \end{cases} \quad (15)$$

where f_g denotes the global text feature.

For each character c_i , we extract its region at the current feature resolution:

$$X_i = \text{Crop}(X, \mathcal{B}_i^X). \quad (16)$$

Let $\Phi_{q,i}^s$, $\Phi_{k,i}^s$, and $\Phi_{v,i}^s$ denote the CI-enhanced self-attention query, key, and value projections for character c_i , respectively, and let $\Phi_{q,i}^c$ denote its enhanced cross-attention query projection:

$$\begin{aligned} Q_i^s &= \Phi_{q,i}^s(X_i), & Q_i^c &= \Phi_{q,i}^c(X_i), \\ K_{i,r}^s &= \Phi_{k,i}^s(X_i), & V_{i,r}^s &= \Phi_{v,i}^s(X_i), \\ K_{i,g}^s &= \Phi_{k,i}^s(X), & V_{i,g}^s &= \Phi_{v,i}^s(X), \end{aligned} \quad (17)$$

where $\Phi_{\cdot,i}$ follows the character-conditioned routing introduced in Sec. 3.2. For self-attention, IA-GD adopts a local-to-global schedule as:

$$O_i^s = \begin{cases} \text{Attn}(Q_i^s, K_{i,r}^s, V_{i,r}^s), & t > T_\theta, \\ \text{Attn}(Q_i^s, K_{i,g}^s, V_{i,g}^s), & t \leq T_\theta, \end{cases} \quad (18)$$

where T_θ controls the transition from regional to global key-value context. During early denoising, restricting the keys and values to the character region strengthens character-specific feature formation. As denoising proceeds, expanding the key-value context to the full feature map allows each character to interact with the surrounding scene, while the region-specific query preserves its identity focus. This transition facilitates more natural character placement and scene composition as shown in Fig.8(b).

For cross-attention, Q_i^c attends to the character-specific text feature f_i , yielding:

$$O_i^c = \text{Attn}(Q_i^c, f_i W_k^c, f_i W_v^c). \quad (19)$$

The character-specific outputs $\{O_i^s, O_i^c\}_{i=1}^m$ replace the corresponding regions in $\{O_g^s, O_g^c\}$, while the global attention responses are retained elsewhere. The resulting attention features are propagated through the full-image denoising forward to produce the full-image noise prediction ϵ_g^t .

IA-GD thus preserves full-image contextual modeling while maintaining identity-aware processing within character regions.

Time-Aware Role-Global Fusion. To balance character identity preservation and global scene coherence, CA-RF dynamically fuses the character-specific noise predictions from RS-CD with the global character-enhanced noise prediction from IA-GD throughout denoising. Let M_i denote the feathered soft mask derived from B_i on the latent canvas. We define:

$$\begin{aligned} S &= \text{clip}\left(\sum_{i=1}^m M_i, 0, 1\right), \\ \omega_i &= \frac{M_i}{\sum_{j=1}^m M_j}, \quad S > 0. \end{aligned} \quad (20)$$

where S denotes the aggregate character-region mask and ω_i denotes the normalized spatial contribution of character c_i . Outside the character regions ($S = 0$), we simply set $\omega_i = 0$. To progressively shift the fusion from character-specific denoising to global denoising, we define the time-aware fusion scale as:

$$\alpha_t = \alpha_{\text{early}} + (\alpha_{\text{late}} - \alpha_{\text{early}}) \pi_t^\gamma, \quad (21)$$

where T denotes the maximum diffusion timestep and $\pi_t = (T-t)/T \in [0, 1]$ denotes the normalized denoising progress, which increases as denoising proceeds. The exponent γ controls the transition rate, and we set $0 \leq \alpha_{\text{early}} < \alpha_{\text{late}} \leq 1$. Accordingly, a smaller α_t emphasizes the role-specific prediction $\hat{\epsilon}_i^t$ during early denoising, whereas its gradual increase shifts the fusion toward the global prediction ϵ_g^t at later steps.

The fused noise prediction is formulated as:

$$\epsilon^t = (1-S) \odot \epsilon_g^t + S \odot \sum_{i=1}^m \omega_i \odot [\alpha_t \epsilon_g^t + (1-\alpha_t) \hat{\epsilon}_i^t]. \quad (22)$$

Outside the character regions, ϵ_g^t is retained unchanged. Within these regions, ω_i determines the spatial contribution of each role-specific prediction, while α_t controls the time-dependent balance between RS-CD and IA-GD. The fused prediction ϵ^t is then used to update z_t .

4 Experiments

4.1 Experimental Settings

4.1.1 Implementation Details.

We implement EvoTale under a continual character-learning protocol and conduct experiments independently in two story worlds, *i.e.*, 9 characters in *Pororo* and 5 characters in *Frozen* from TBC-Bench(Zhang et al. 2025). We use SD1.5 with the Rev-Animated checkpoint¹ as the default backbone. We additionally report EvoTale with SDXL (Podell et al. 2024) using the Starlight checkpoint². Methods (Labs et al. 2025; Wang et al. 2025; He et al. 2025a) based on SDXL or DiT (Peebles and Xie 2023) are included as cross-backbone references.

For each character c_j , we optimize L_j and the layer-wise character tokens by minimizing the standard diffusion denoising MSE loss. Optimization is performed using AdamW with an initial learning rate of 5×10^{-4} . For the initial customization attempt, each character-specific training set is repeated 500 times, so that the initial training budget scales with the number of reference images. We set $r = 80$ and $r_1 = 20$. For **CQG**, we use Qwen3-VL-4B-Instruct as the MLLM and set $K = 5$, $\tau = 3$ on a 0–4 scale, $M = 8$, $n = 3$ independently sampled reference images per complete evaluation, and $\rho_{\min} = 0.2$. For **CA-RF**, we set $T_\theta = 780$ and use staged global prompts, specifying the scene early and introducing character actions and interactions later. We use $(\alpha_{\text{early}}, \alpha_{\text{late}}, \gamma) = (0.2, 0.8, 2.0)$ for SD1.5 and $(0.2, 0.6, 2.0)$ for SDXL. For methods requiring spatial priors, we follow their official default configurations. For EvoTale, coarse character bounding boxes are arranged approximately evenly from left to right according to prompt order and remain fixed during sampling. All experiments are conducted on 4 GeForce RTX 4090 GPUs.

4.1.2 Comparison Methods.

We benchmark EvoTale against four categories of comparison methods for single- and multi-character story visualization. These include tuning-free customization methods (Labs et al.

2025; He et al. 2025b), tuning-based customization and composition methods (Gu et al. 2023; Yang et al. 2025; Zhong et al. 2024; Po et al. 2024), sequential or continual customization methods (Gal et al. 2023; Dong et al. 2024), and different story visualization methods (Zhang et al. 2025; He et al. 2025a; Wang et al. 2025). We categorize Textual Inversion (Gal et al. 2023) as a continual method because character-specific token embeddings can be learned sequentially without modifying previously acquired embeddings. Tuning-free methods are evaluated using their official implementations. For optimization-based baselines implemented in our experiments, we use the same character images and base model checkpoint as EvoTale.

4.1.3 Evaluation Metrics.

We evaluate character alignment and text–image semantic alignment for both settings. Character alignment is measured using full-frame DINO similarity (**D-I**) (Oquab et al. 2024) and CLIP image similarity (**C-I**) (Radford et al. 2021) between each generated frame and the corresponding character reference images, while semantic alignment is measured using CLIP text–image similarity (**C-T**) between each frame and its corresponding story description. In the single-character setting, D-I and C-I are computed against the target character references. In the multi-character setting, they are computed against the references of each character appearing in the prompt and then averaged across characters. For each story world, the single-character setting evaluates each character on five stories, each of which contains only that character, whereas the multi-character setting uses five stories involving multiple characters. We generate five sequences per story and average all metrics over the resulting frames.

4.2 Single-Character Comparisons

4.2.1 Quantitative Experiments.

As shown in Tab. 1, DisEnvisioner achieves the highest C-I on *Frozen*, but its relatively low C-T indicates weaker adaptation to story semantics. Its lower identity scores on *Pororo* further suggest difficulty in capturing the distinctive attributes of highly stylized characters. In contrast, Textual

¹<https://civitai.com/models/7371/rev-animated>

²<https://huggingface.co/stablediffusionapi/starlight-xl-v3/tree/main>

Table 1 Quantitative comparisons on single-character story settings. The upper block reports comparisons among SD1.5-based methods, while the lower block provides cross-backbone reference results. Best results are shown in **bold** and the second-best are underlined. * denotes DiT-based models and † denotes SDXL-based models; unmarked methods use SD1.5.

Type	Method	Pororo			Frozen		
		D-I↑	C-I↑	C-T↑	D-I↑	C-I↑	C-T↑
SD1.5-Based Comparison							
Tuning-free	DisEnvisioner	47.03	74.76	34.05	<u>49.63</u>	83.53	32.80
Tuning-based	Mix-of-Show	49.89	79.38	34.17	45.05	80.66	35.08
Tuning-based	Orth-Adapt	<u>52.01</u>	<u>79.99</u>	34.06	49.61	82.56	35.10
CL	Textual Inversion	42.99	73.32	34.92	36.40	73.40	35.46
CL	CIDM	47.91	78.67	32.65	44.06	80.85	32.59
CL	Ours	54.46	80.89	<u>34.42</u>	51.50	<u>82.84</u>	<u>35.23</u>
Cross-Backbone References							
Story visualization(Training-free)	CharaConsist*	49.52	74.70	34.68	37.92	80.76	34.88
Tuning-free	FLUX.1 Kontext*	55.06	<u>79.03</u>	35.19	<u>45.08</u>	79.88	36.24
CL	Ours [†]	<u>54.67</u>	82.23	<u>35.04</u>	52.78	85.67	<u>35.64</u>

Inversion achieves the highest C-T, but its substantially lower D-I and C-I reveal limited preservation of fine-grained character appearance. Other tuning-based and CL methods generally show stronger identity preservation. For example, Orth-Adapt provides more balanced performance and is particularly competitive on *Pororo*, but requires post-hoc merging of separately learned character adaptations. EvoTale achieves a stronger overall balance without such model composition, obtaining the highest D-I in SD1.5 comparison. It improves over the strongest competing method by +4.71% on *Pororo* and +3.77% on *Frozen*, while maintaining competitive C-I and C-T.

Among the cross-backbone references, CharaConsist shows weaker reference-identity alignment due to its focus on cross-image consistency rather than explicit reference-appearance modeling. FLUX.1 Kontext achieves the highest D-I on *Pororo*. However, as shown in Figs. 1 and 4, this result partly reflects overly literal preservation of the reference image with limited story-conditioned adaptation. Ours maintains strong identity preservation on both datasets with favorable C-T, indicating a better balance between reference fidelity and story adaptation.

4.2.2 Qualitative Experiments.

We additionally include StoryWeaver and DreamStory, whose outputs use different aspect ratios,

for qualitative comparisons in Fig. 4. The qualitative results further explain the performance differences observed in Tab. 1. DisEnvisioner often produces portrait-like compositions and retains reference-specific poses. For example, its outputs for V6 in *Pororo* and U5 in *Frozen* remain close to the reference viewpoint and pose, despite being conditioned on new story contexts. Such strong adherence to the reference helps preserve character similarity but limits adaptation to diverse story contexts. On *Pororo*, it also captures distinctive character attributes less precisely, consistent with its weaker identity scores. Textual Inversion better follows the requested scenes but often loses fine-grained appearance details. FLUX.1 Kontext also shows strong scene alignment, yet occasionally preserves reference-specific cues too literally. For example, V2 retains the roller skates from the reference despite the absence of any skating-related context in the story prompt, with the main changes concentrated in the surrounding scene. DreamStory and CharaConsist primarily emphasize maintaining an established character appearance across generated images rather than explicitly matching the user-provided reference character, which can lead to identity deviations in some cases. Other tuning-based and continual customization methods generally improve character fidelity but may still exhibit appearance drift across scenes and poses. In comparison, EvoTale more consistently preserves fine-grained character

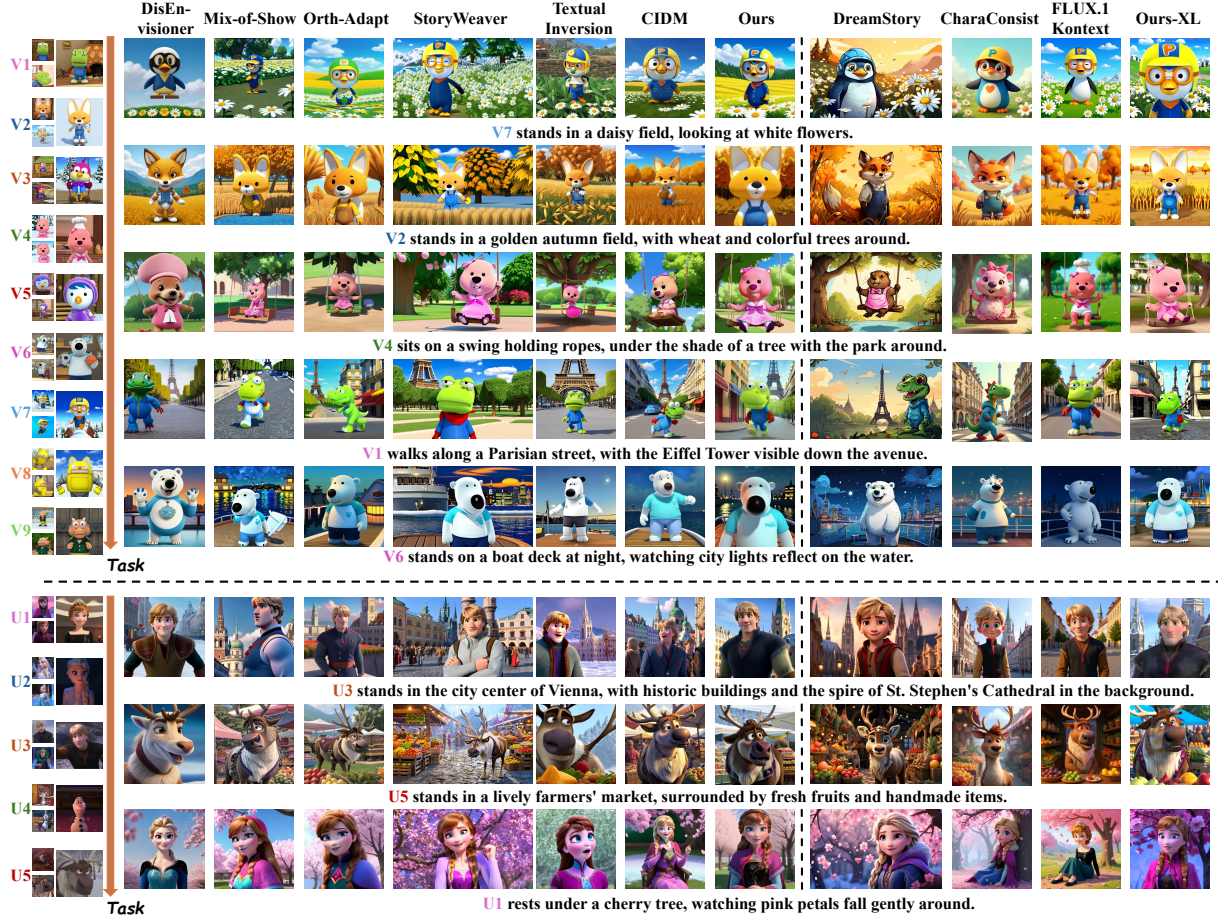


Fig. 4 Qualitative comparison of single-character story visualization. The arrows indicate the character learning order for continual-learning methods. EvoTale preserves character identity while maintaining favorable alignment with the story descriptions across both story settings.

identities while naturally adapting them to diverse story scenes, poses, and actions.

4.3 Character Identity Preservation

To further assess character identity preservation during story-world expansion, we conduct two complementary analyses, *i.e.*, identity retention across sequential updates and stability under multi-character integration.

4.3.1 Continual Identity Preservation.

We evaluate accumulated forgetting by tracking the first learned character, which experiences the longest sequence of subsequent updates, as shown in Fig. 5. We compare EvoTale with CIDM and two sequential LoRA baselines: *LoRA-Tuning*,

which sequentially updates a shared LoRA module, and *LoRA-Tuning+LT*, which additionally learns a character-specific token embedding. Since these methods differ in their adaptation mechanisms and capacity allocation, we focus on identity stability across learning stages rather than absolute scores at individual stages. As shown in Fig. 5(a) and Fig. 5(b), across both *Pororo* and *Frozen*, the two sequential LoRA baselines exhibit noticeable fluctuations and an overall decline in D-I and C-I as new characters are introduced. This behavior indicates accumulated interference from subsequent updates. CIDM is comparatively stable, but its identity scores still gradually decrease along the learning stream. In contrast, EvoTale maintains stable D-I and C-I scores with only minor variations across stages. The qualitative

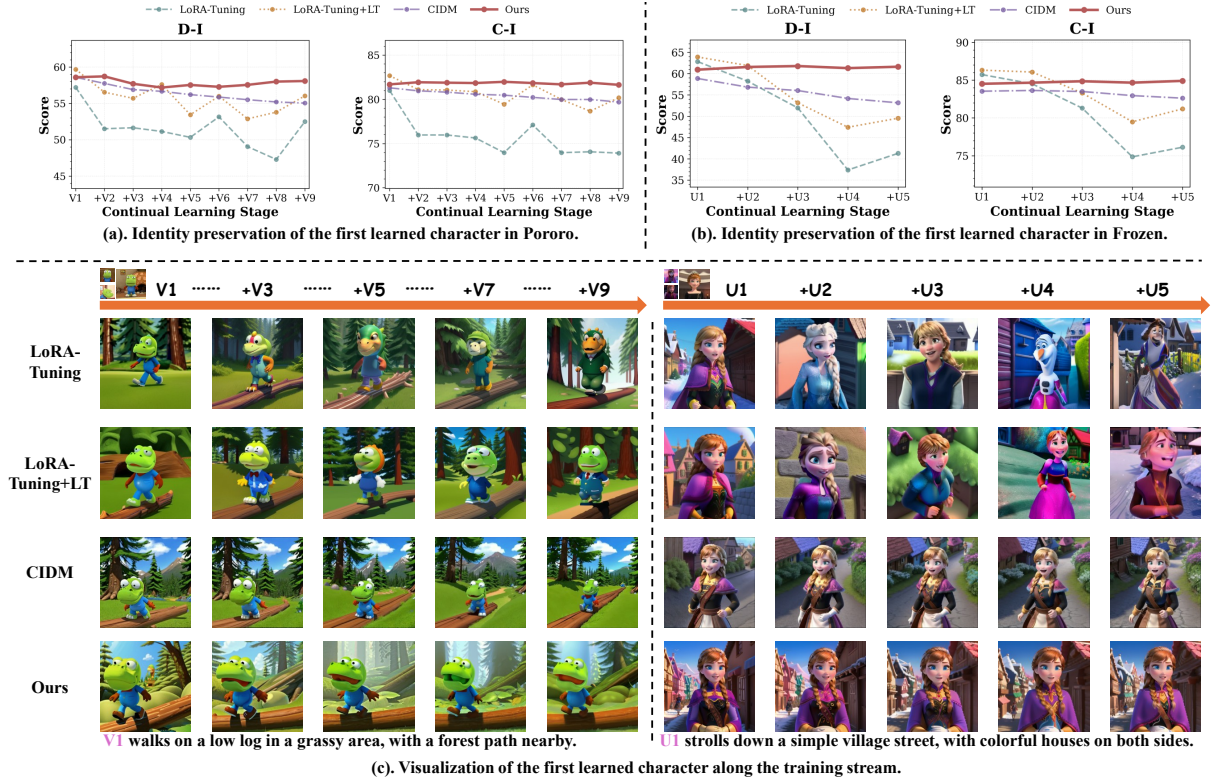


Fig. 5 Identity preservation during continual integration. (a)–(b) Identity scores of the first learned character across learning stages on *Pororo* and *Frozen*. (c) Qualitative results across stages, with arrows indicating the learning order.

Table 2 Before-and-after comparison of character customization performance. For Mix-of-Show and Orth-Adapt, *Before* denotes independently trained character modules, while *After* denotes the merged model. For EvoTale, *Before* denotes the average performance of each character immediately after its introduction, while *After* denotes its average performance in the final continually updated model. Δ denotes *After* minus *Before*.

Dataset	Method	D-I $\uparrow$			C-I $\uparrow$			C-T $\uparrow$		
		Before	After	Δ	Before	After	Δ	Before	After	Δ
<i>Pororo</i>	Mix-of-Show	53.40	49.89	-3.51	81.13	79.38	-1.75	32.71	34.17	+1.46
	Orth-Adapt	54.18	52.01	-2.17	80.76	79.99	-0.77	33.60	34.06	+0.46
	EvoTale	54.85	54.46	-0.39	80.92	80.89	-0.03	34.73	34.42	-0.31
<i>Frozen</i>	Mix-of-Show	49.99	45.05	-4.94	83.26	80.66	-2.60	34.46	35.08	+0.62
	Orth-Adapt	51.96	49.61	-2.35	84.51	82.56	-1.95	34.65	35.10	+0.45
	EvoTale	52.31	51.50	-0.81	83.12	82.84	-0.28	35.42	35.23	-0.19

comparisons in Fig. 5(c) further show that sequential LoRA updates may alter defining character attributes or cause identity drift toward newly learned characters, whereas EvoTale more consistently preserves the first character throughout the continual learning stream.

4.3.2 Character Customization Stability.

We further examine whether learned character identities remain stable during multi-character integration. As shown in Tab. 2, both composition-based methods exhibit identity

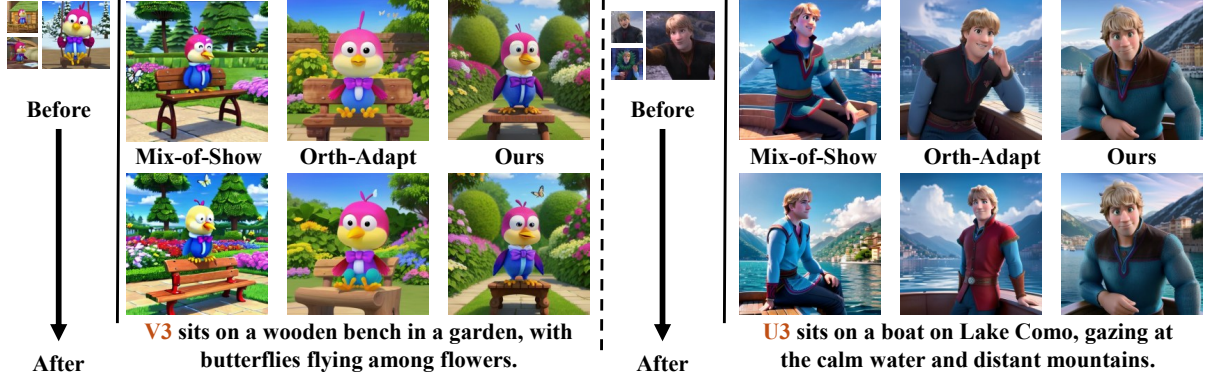


Fig. 6 Character identity before and after multi-character integration. EvoTale exhibits smaller identity changes than post-hoc composition methods after integrating multiple characters.

degradation after model composition, with Orth-Adapt showing better stability. In comparison, EvoTale shows substantially smaller before-to-after changes across both story worlds. The qualitative examples in Fig. 6 show a similar trend. For example, Mix-of-Show changes V3’s head coloration and U3’s facial appearance, while Orth-Adapt changes V3’s clothing. In contrast, EvoTale better preserves character identity as more characters are integrated. These results indicate that organizing character-specific knowledge within a unified LoRA branch through orthogonal-basis-guided routing enables more stable story-world expansion without post-hoc module composition.

4.4 Multi-Character Comparisons

4.4.1 Quantitative Experiments.

Quantitative results for multi-character visual storytelling are shown in Tab. 3. LoRA-S alternates among character-specific LoRAs during denoising, whereas LoRA-C aggregates their predictions at each step. Without explicit character-specific spatial routing, both methods remain susceptible to identity confusion in multi-character scenes, as reflected by their relatively low D-I scores on *Pororo*. Mix-of-Show combines customized concepts through post-hoc gradient fusion and spatial constraints. When pose conditioning is disabled in Mix-of-Show*, its lower D-I indicates weaker identity preservation. Additionally, merging five characters also takes about 11 minutes in our setting. LoRA-Composer assigns individual LoRAs to designated regions, but its latent re-initialization and

constraint-guided refinement increase generation time to about 60s per image. StoryWeaver preserves identity through knowledge-enhanced spatial guidance to inject character semantics, yet its implicit spatial guidance remains limited in balancing character fidelity and new story semantics.

In comparison, EvoTale achieves a better balance between character identity and story-text alignment. Under SD1.5, it maintains competitive identity preservation with strong story-text alignment on *Pororo*, and particularly strong identity preservation on *Frozen* without sacrificing semantic alignment. It requires about 30s per image without post-hoc model fusion or iterative latent refinement. Under SDXL, DreamStory shows favorable story-text alignment but weaker reference-identity preservation and requires about 88s per image, whereas EvoTale-SDXL takes about 68s while achieving substantially higher D-I and C-I with comparable semantic alignment. All inference times are measured on a single GeForce RTX 4090 GPU.

4.4.2 Qualitative Experiments.

As shown in Fig. 7, existing methods exhibit different limitations in multi-character story visualization. LoRA-S and LoRA-C tend to omit required characters or to exhibit identity blending among characters within the same frame. The key-pose guidance in Mix-of-Show helps maintain character placement but can result in more constrained compositions. Without pose guidance, Mix-of-Show* allows greater layout flexibility but becomes more

Table 3 Quantitative comparison on multi-character story visualization. The upper block compares methods in SD1.5, while the lower block reports SDXL reference results. Best results are shown in **bold** and the second-best are underlined, while only the best result is highlighted in the SDXL block. † denotes SDXL-based models; unmarked methods use SD1.5.

Type	Method	Pororo			Frozen		
		D-I↑	C-I↑	C-T↑	D-I↑	C-I↑	C-T↑
SD1.5-Based Comparison							
Tuning-based	LoRA-S	37.24	65.49	33.60	37.47	72.13	33.90
	LoRA-C	37.31	65.54	33.85	42.38	74.20	34.28
	Mix-of-Show	40.27	69.13	34.72	<u>43.57</u>	74.34	36.11
	Mix-of-Show*	36.23	68.04	34.56	39.24	75.49	36.08
	LoRA-Composer	40.00	67.69	<u>36.56</u>	41.35	<u>76.18</u>	37.17
Story visualization(Tuning-based)	StoryWeaver	53.38	72.61	34.03	42.19	75.98	36.15
CL	Ours	<u>44.95</u>	<u>69.54</u>	36.59	45.32	77.98	<u>36.44</u>
SDXL Reference							
Story visualization(Training-free)	DreamStory [†]	31.50	63.39	37.14	39.68	73.11	36.55
CL	Ours [†]	46.45	71.04	36.66	46.35	79.20	36.78

* For Mix-of-Show*, the pose-conditioning weight is set to zero while retaining the key-pose adapter computation.



Fig. 7 Qualitative comparison of multi-character story visualization. EvoTale better preserves distinct character identities while maintaining natural interactions and strong alignment with the story descriptions.

susceptible to character omission or identity deviation. LoRA-Composer and StoryWeaver generally render both characters within coherent scenes, although identity deviations remain in some scenes. For example, LoRA-Composer loses some distinctive appearance details of V5 in the 2nd example, while the visual traits of U1 and U2 become

partially entangled in StoryWeaver’s tea-party result in the 4th row. DreamStory follows the requested scenes and character interactions well, but tends to generate semantically plausible characters rather than faithfully preserving the specified identities. In comparison, our method more consistently retains the distinct identities of the

Table 4 User study on single- and multi-character story visualization. C-A, S-A, and V-Q denote character alignment, story alignment, and visual quality, respectively. Best results are shown in **bold** and the second-best are underlined. * denotes DiT-based models and † denotes SDXL-based models; unmarked methods use SD1.5.

Method	C-A↑	S-A↑	V-Q↑
<i>Single Character</i>			
DreamStory [†]	1.56	3.52	3.80
CharaConsist*	1.80	3.44	3.56
DisEnvisioner	1.44	2.36	2.12
FLUX.1 Kontext*	3.12	<u>3.76</u>	3.28
Mix-of-Show	2.64	2.88	2.92
Orth-Adapt	3.20	3.04	3.12
StoryWeaver	3.96	3.44	3.60
Textual Inversion	2.44	2.88	2.32
CIDM	2.48	2.84	2.60
Ours	<u>4.24</u>	3.68	<u>3.84</u>
Ours-XL [†]	4.48	3.84	4.08
<i>Multi-Character</i>			
DreamStory [†]	1.52	<u>4.12</u>	3.68
LoRA-S	1.32	1.24	2.72
LoRA-C	1.40	1.84	2.80
Mix-of-Show	2.44	2.92	2.76
Mix-of-Show*	2.32	2.68	2.60
LoRA-Composer	2.72	2.92	3.16
StoryWeaver	3.72	3.64	3.56
Ours	<u>4.08</u>	3.92	4.20
Ours-XL [†]	4.28	4.16	<u>4.12</u>

prompted characters while depicting the requested scenes and interactions. Overall, EvoTale achieves a favorable balance between multi-character identity fidelity and story-text alignment.

4.5 User Study

We further conduct a user study with 25 participants to assess the perceptual quality of different story visualization methods. Participants assign ratings for each method based on anonymously presented generated sequences, considering Character Alignment (C-A), Story Alignment (S-A), and Overall Visual Quality (V-Q) on a five-point Likert scale, with higher scores indicating better performance. As shown in Tab. 4, EvoTale receives consistently high ratings across the three criteria, indicating favorable human evaluations of EvoTale in terms of character alignment, story alignment, and overall visual quality.

4.6 Ablation Study

4.6.1 Quantitative Analysis

Single-character setting. We first evaluate CI and CQG in Tab. 5. To ablate CI, we use a single shared LoRA branch that is sequentially updated for all characters without replaying data from previously learned characters. All other training settings are kept unchanged. This baseline decreases D-I and C-I on both *Pororo* and *Frozen*, demonstrating that CI’s character-specific residual accumulation and routing are important for limiting cross-character interference and preserving previously learned identities. To evaluate CQG, we compare it with two fixed schedules representing relatively conservative and intensive customization budgets. Specifically, each character-specific training set is repeated 500 and 1,500 times under the base and large budgets, respectively. Without CQG, the fixed base budget yields lower identity scores than the full model, indicating insufficient adaptation for some characters. Increasing the fixed budget improves D-I and C-I but noticeably degrades C-T, suggesting that more intensive optimization increasingly favors reference appearance over story-text alignment. By adjusting the refinement budget according to the assessed customization quality, CQG achieves a more favorable balance between identity preservation and semantic alignment.

The lower block examines the parameterization of CI. For a fair comparison, standard LoRA and all CI variants are applied to the same layers and trained with the same base budget. With the shared factor B fixed, CI without learnable tokens obtains lower D-I and C-I than standard LoRA on both story worlds, indicating reduced flexibility in character-specific adaptation. Character-specific learnable token embeddings compensate for this constraint by providing an additional pathway for identity representation. With these embeddings, CI achieves higher D-I than standard LoRA on both story worlds and higher C-I on *Pororo*, while remaining close on *Frozen*.

Multi-character setting. Tab. 6 evaluates the key components of Character-Aware Region-Focus Sampling. Compared with fixed fusion, time-aware fusion between RS-CD and standard global denoising decreases identity preservation.

Table 5 Ablation study of the main components and CI parameterization on single-character story visualization. For each dataset, best results are shown in **bold** and the second-best are underlined. LT denotes learnable token embeddings.

Setting	Pororo			Frozen		
	D-I↑	C-I↑	C-T↑	D-I↑	C-I↑	C-T↑
<i>Effect of Main Components</i>						
Shared LoRA (Sequential, No Replay)	45.75	77.04	35.61	42.84	78.75	36.08
w/o CQG (Fixed Base Budget)	51.85	79.53	<u>35.45</u>	47.63	80.88	<u>35.73</u>
w/o CQG (Fixed Large Budget)	59.32	83.84	31.96	56.42	85.03	33.19
Full Model	<u>54.46</u>	<u>80.89</u>	34.42	<u>51.50</u>	<u>82.84</u>	35.23
<i>Effect of CI Parameterization</i>						
LoRA (Fixed Base Budget)	<u>47.28</u>	<u>77.06</u>	<u>35.60</u>	<u>47.02</u>	81.62	35.53
CI w/o LT (Fixed Base Budget)	46.88	74.78	35.91	46.28	79.84	36.11
CI w/ LT (Fixed Base Budget)	51.85	79.53	35.45	47.63	<u>80.88</u>	<u>35.73</u>

Table 6 Ablation study of CA-RF on multi-character story visualization. Best results are shown in **bold** and the second-best are underlined. RS-CD and IA-GD denote Role-Specific Crop Denoising and Identity-Aware Global Denoising, and Global denotes standard global denoising. Cross-AR and Regional-SA denote character-specific cross-attention routing and region-restricted self-attention, respectively.

Setting	Pororo			Frozen		
	D-I↑	C-I↑	C-T↑	D-I↑	C-I↑	C-T↑
RS-CD + Global (Fixed Fusion)	42.87	68.18	36.43	43.32	76.19	36.05
RS-CD + Global (Time-Aware Fusion)	37.29	65.95	36.31	38.85	75.11	35.54
RS-CD + IA-GD (Cross-AR)	42.68	68.13	36.79	42.53	<u>77.16</u>	36.04
RS-CD + IA-GD (Cross-AR + Regional-SA)	46.35	69.73	36.25	45.65	76.97	36.57
Full CA-RF	<u>44.95</u>	<u>69.54</u>	<u>36.59</u>	<u>45.32</u>	77.98	<u>36.44</u>

Replacing the standard global branch with IA-GD using Cross-AR recovers both identity and semantic alignment, while further restricting self-attention to character regions improves D-I on both story worlds. Full CA-RF introduces a local-to-global self-attention transition, allowing character features to interact with the surrounding scene as denoising proceeds. Although keeping self-attention region-restricted throughout denoising yields slightly higher scores on several metrics, the qualitative results in Fig.8(b) show that it is more susceptible to rigid character layouts. Considering both quantitative performance and scene composition, we adopt full CA-RF as the final sampling strategy.

4.6.2 Qualitative Analysis

As shown in Fig. 8(a), using a single shared LoRA branch causes visible identity drift and loss of fine-grained attributes. Without CQG, the fixed

base budget may provide insufficient adaptation for some characters, whereas the fixed large budget better reproduces character appearance but can weaken scene details and semantic alignment. For example, the fixed large budget preserves the character’s appearance well but fails to generate “meadow filled with wildflowers” described in the story prompt for V7. Regarding CI parameterization, fixing the shared factor B without learnable tokens limits character-specific adaptation and leads to greater loss of reference-specific identity details than standard LoRA. Introducing character-specific learnable tokens into CI compensates for this constraint and recovers finer identity cues. In contrast, the full setting better balances identity preservation and story alignment, showing the complementary effects of character-specific subspace assignment, learnable tokens, and adaptive budget allocation.

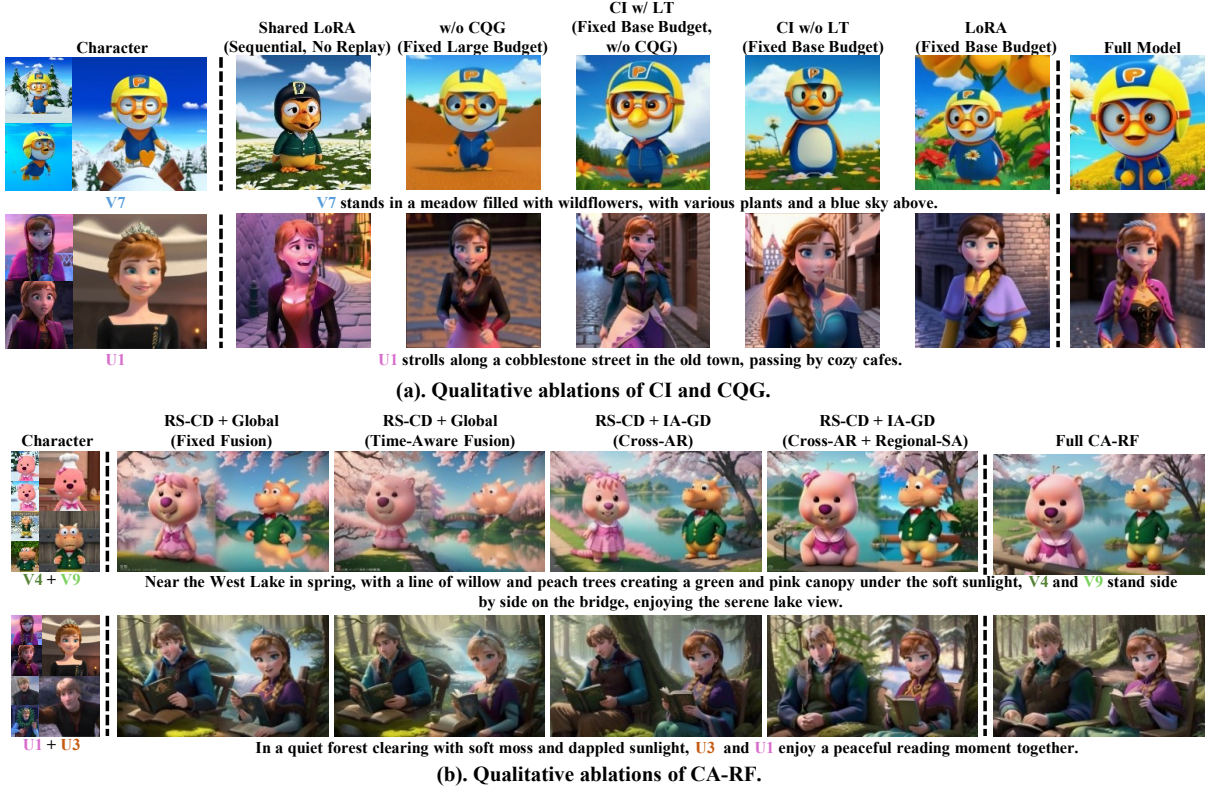


Fig. 8 Qualitative ablation results for the key components of EvoTale. (a) CI and CQG in single-character story settings. (b) CA-RF sampling in multi-character story settings.

Results in Fig. 8(b) further illustrate the effects of different CA-RF components. Fixed fusion between RS-CD and standard global denoising provides strong regional identity guidance but tends to produce relatively rigid character layouts. Introducing time-aware fusion increases the contribution of global denoising and improves composition flexibility, but identity preservation becomes weaker when the global branch lacks identity-aware routing. Replacing standard global denoising with IA-GD using Cross-AR improves character assignment and global scene coordination. Further restricting self-attention within character regions strengthens identity preservation and produces more natural compositions than fixed fusion. However, the persistent regional separation can introduce visible seam-like boundaries between character regions, as observed in both examples in Fig. 8(b). By allowing character features to progressively interact with the surrounding scene during denoising,

Full CA-RF reduces such region-wise discontinuities and yields more spatially coherent compositions while preserving the prompted identities. These observations are consistent with the quantitative results in Tab. 6.

5 Conclusion

In this paper, we present EvoTale, a continual character customization framework for expanding stylized story worlds. We introduce the All-in-One-World Character Integrator to accumulate character-specific residual components within a unified LoRA branch through orthogonal-basis-guided routing, while freezing previously learned components to limit cross-character interference. We further develop the Character Quality Gate to adaptively control the optimization budget through rubric-guided MLLM assessment, and Character-Aware Region-Focus Sampling to combine regional identity guidance with identity-aware global denoising for coherent

multi-character generation. Experimental results demonstrate a favorable balance across character fidelity, story-text alignment, continual identity retention, and multi-character generation quality.

References

- Bai S, Cai Y, Chen R, et al (2025) Qwen3-vl technical report. arXiv preprint arXiv:251121631
- Banerjee A, Lladós J, Pal U, et al (2025) Talediffusion: Multi-character story generation with dialogue rendering. arXiv preprint arXiv:250904123
- Bugliarello E, Moraldo HH, Villegas R, et al (2023) Storybench: A multifaceted benchmark for continuous story visualization. *Advances in Neural Information Processing Systems* 36:78095–78125
- Cha H, Lee J, Shin J (2021) Co2l: Contrastive continual learning. In: *Proceedings of the IEEE/CVF International conference on computer vision*, pp 9516–9525
- Chen H, Han R, Wu TL, et al (2022) Character-centric story visualization via visual planning and token alignment. In: Goldberg Y, Kozareva Z, Zhang Y (eds) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp 8259–8272, <https://doi.org/10.18653/v1/2022.emnlp-main.565>
- Cotogni M, Yang F, Cusano C, et al (2025) Exemplar-free continual learning of vision transformers via gated class-attention and cascaded feature drift compensation. *International Journal of Computer Vision* 133(7):4571–4589
- Dong J, Liang W, Li H, et al (2024) How to continually adapt text-to-image diffusion models for flexible customization? *Advances in Neural Information Processing Systems* 37:130057–130083
- Dong R, Pang W, Pan C, et al (2025) Storycrafter: Instance-aligned multi-character storytelling with diffusion policy learning. In: *Proceedings of the 33rd ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, MM ’25, p 3232–3241, <https://doi.org/10.1145/3746027.3755022>
- Gal R, Alaluf Y, Atzmon Y, et al (2023) An image is worth one word: Personalizing text-to-image generation using textual inversion. In: *The Eleventh International Conference on Learning Representations*
- Gu Y, Wang X, Wu JZ, et al (2023) Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in neural information processing systems* 36:15890–15902
- Guo Z, Jin T (2025) Conceptguard: Continual personalized text-to-image generation with forgetting and confusion mitigation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 2945–2954
- Han L, Li Y, Zhang H, et al (2023) Svdiff: Compact parameter space for diffusion fine-tuning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 7323–7334
- He H, Yang H, Tuo Z, et al (2025a) Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47(12):11874–11891. <https://doi.org/10.1109/TPAMI.2025.3600149>
- He J, Li H, Hu Y, et al (2025b) Disenvisioner: Disentangled and enriched visual prompt for customized image generation. In: Yue Y, Garg A, Peng N, et al (eds) *International Conference on Learning Representations*, pp 80652–80664
- Hoang TV, Nguyen QB, Do TT, et al (2025) Showflow: From robust single concept to condition-free multi-concept generation. arXiv preprint arXiv:250618493

- Hu EJ, Shen Y, Wallis P, et al (2021) Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:210609685
- Huang Y, Qin Y, Lu S, et al (2024) Story3d-agent: Exploring 3d storytelling visualization with large language models. arXiv preprint arXiv:240811801
- Jeong H, Kwon G, Ye JC (2023) Zero-shot generation of coherent storybook from plain text story using diffusion models. arXiv preprint arXiv:230203900
- Jiang J, Zhang Y, Feng K, et al (2025) Mc2: Multi-concept guidance for customized multi-concept generation. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 2802–2812, <https://doi.org/10.1109/CVPR52734.2025.00267>
- Jin J, Yu Z, Shen Y, et al (2025) Latexblend: Scaling multi-concept customized generation with latent textual blending. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 23585–23594, <https://doi.org/10.1109/CVPR52734.2025.02196>
- Kong Y, Liu L, Qiao M, et al (2023) Trust-region adaptive frequency for online continual learning. *International Journal of Computer Vision* 131(7):1825–1839
- Kong Z, Zhang Y, Yang T, et al (2025) Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In: Leonardis A, Ricci E, Roth S, et al (eds) *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, pp 253–270
- Kumari N, Zhang B, Zhang R, et al (2023) Multi-concept customization of text-to-image diffusion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1931–1941
- Kwon G, Jenni S, Li D, et al (2024) Concept weaver: Enabling multi-concept fusion in text-to-image models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 8880–8889
- Labs BF, Batifol S, Blattmann A, et al (2025) Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:250615742
- Li B (2022) Word-level fine-grained story visualization. In: Avidan S, Brostow G, Cissé M, et al (eds) *Computer Vision – ECCV 2022*. Springer Nature Switzerland, Cham, pp 347–362
- Li Y, Gan Z, Shen Y, et al (2019) Storygan: A sequential conditional gan for story visualization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*
- Liu C, Wu H, Zhong Y, et al (2024) Intelligent grimm - open-ended visual storytelling via latent diffusion models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 6190–6200, <https://doi.org/10.1109/CVPR52733.2024.00592>
- Ma KL, Liao I, Frazier J, et al (2012) Scientific storytelling using visualization. *IEEE Computer Graphics and Applications* 32(1):12–19. <https://doi.org/10.1109/MCG.2012.24>
- Oquab M, Darcet T, Moutakanni T, et al (2024) DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research* URL <https://openreview.net/forum?id=a68SUt6zFt>
- Pan X, Qin P, Li Y, et al (2024) Synthesizing coherent story with auto-regressive latent diffusion models. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 2908–2918, <https://doi.org/10.1109/WACV57701.2024.00290>
- Peebles W, Xie S (2023) Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp 4195–4205
- Peng Y, Cui Y, Tang H, et al (2025) Dreambench++: A human-aligned benchmark for personalized image generation. In: Yue Y, Garg A, Peng N, et al (eds) *International Conference on Learning Representations*, pp 46010–46032

- Po R, Yang G, Aberman K, et al (2024) Orthogonal adaptation for modular customization of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 7964–7973
- Podell D, English Z, Lacey K, et al (2024) Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations, pp 1862–1874
- Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PmLR, pp 8748–8763
- Ruiz N, Li Y, Jampani V, et al (2023) 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 (CVPR), pp 22500–22510
- Shen X, Elhoseiny M (2025) Storygpt-v: Large language models as consistent story visualizers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 13273–13283
- Smith JS, Hsu YC, Zhang L, et al (2023) Continual diffusion: Continual customization of text-to-image diffusion with c-lora. arXiv preprint arXiv:230406027
- Song YZ, Rui Tam Z, Chen HJ, et al (2020) Character-preserving coherent story visualization. In: Vedaldi A, Bischof H, Brox T, et al (eds) Computer Vision – ECCV 2020. Springer International Publishing, Cham, pp 18–33
- Tao M, Bao BK, Tang H, et al (2025) Storyimager: A unified and efficient framework for coherent story visualization and completion. In: Leonardi A, Ricci E, Roth S, et al (eds) Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp 479–495
- Tong C, Roberts R, Borgo R, et al (2018) Storytelling and visualization: An extended survey. Information 9(3). <https://doi.org/10.3390/info9030065>, URL <https://www.mdpi.com/2078-2489/9/3/65>
- Wang L, Zhang X, Su H, et al (2024a) A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence 46(8):5362–5383
- Wang M, Ding H, Peng J, et al (2025) Characonsistent: Fine-grained consistent character generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 16058–16067
- Wang Q, Bai X, Wang H, et al (2024b) Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:240107519
- Wang X, Wang Y, Tsutsui S, et al (2024c) Evolving storytelling: Benchmarks and methods for new character customization with diffusion models. In: Proceedings of the 32nd ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, MM '24, p 3751–3760, <https://doi.org/10.1145/3664647.3681373>
- Wang Z, Zhang Z, Lee CY, et al (2022) Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 139–149
- Wei Y, Zhang Y, Ji Z, et al (2023) Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 15943–15953
- Xiao G, Yin T, Freeman WT, et al (2025) Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision 133(3):1175–1194
- Yang Y, Wang W, Peng L, et al (2025) Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models. IEEE Transactions on Image Processing 34:8145–8158. <https://doi.org/10.1109/TIP.2025.3633153>

- Ye H, Zhang J, Liu S, et al (2023) Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:230806721
- Zhang J, Tang J, Zhang R, et al (2025) Storyweaver: A unified world model for knowledge-enhanced story character customization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 9951–9959
- Zhao C, Ding G, Wang W, et al (2026) Freercustom: Training-free multi-concept customization for image and video generation. *International Journal of Computer Vision* 134(1):17
- Zheng S, Fu Y (2025) Contextualstory: Consistent visual storytelling with spatially-enhanced and storyline context. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 10617–10625
- Zhong M, Shen Y, Wang S, et al (2024) Multilora composition for image generation. arXiv preprint arXiv:240216843
- Zhou T, Duan Z, Chen C, et al (2025) Agentstory: A multi-agent system for story visualization with multi-subject consistent text-to-image generation. In: Proceedings of the 2025 International Conference on Multimedia Retrieval. Association for Computing Machinery, New York, NY, USA, ICMR '25, p 1894–1902, <https://doi.org/10.1145/3731715.3733271>
- Zhou Y, Zhou D, Cheng MM, et al (2024) Storydiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems* 37:110315–110340
- Zhu J, Ma H, Chen J, et al (2025) Domainstudio: Fine-tuning diffusion models for domain-driven image generation using limited data. *Int J Comput Vision* 133(10):7012–7036. <https://doi.org/10.1007/s11263-025-02498-0>, URL <https://doi.org/10.1007/s11263-025-02498-0>
- Zhuang C, Huang A, Hu Y, et al (2026) Vistorybench: Comprehensive benchmark suite for story visualization. In: Proceedings of the IEEE/CVF Conference on Computer Vision