StoryWeaver: A Unified World Model for Knowledge-Enhanced Story Character Customization

Jinlu Zhang[1*], **Jiji Tang**[2*], **Rongsheng Zhang**[2], **Tangjie Lv**[2], **Xiaoshuai Sun**[1†]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
[2]Fuxi AI Lab, Netease Inc., Hangzhou, China
zhangjinlu@stu.xmu.edu.cn, {tangjiji01, zhangrongsheng, hzlvtangjie}@corp.netease.com, xssun@xmu.edu.cn

Abstract

Story visualization has gained increasing attention in artificial intelligence. However, existing methods still struggle with maintaining a balance between character identity preservation and text-semantics alignment, largely due to a lack of detailed semantic modeling of the story scene. To tackle this challenge, we propose a novel knowledge graph, namely Character-Graph (CG), which represents various story-related knowledge, including the characters, their attributes and the relationship. We then introduce StoryWeaver, an image generator that achieves Customization via Character-Graph (C-CG), capable of consistent story visualization with rich text semantics. To further improve the multi-character generation performance, we incorporate knowledge-enhanced spatial guidance (KE-SG) into StoryWeaver to precisely inject character semantics into generation. To validate the effectiveness of our proposed method, extensive experiments are conducted using a new benchmark called TBC-Bench. The experiments confirm that our StoryWeaver excels not only in creating vivid visual story plots but also in accurately conveying character identities across various scenarios with considerable storage efficiency, e.g., achieving an average increase of +9.03% DINO-I and +13.44% CLIP-T. Furthermore, ablation experiments are conducted to verify the superiority of each proposed module.

Code — https://github.com/Aria-Zhangjl/StoryWeaver
Extended version — https://arxiv.org/abs/2412.07375

Introduction

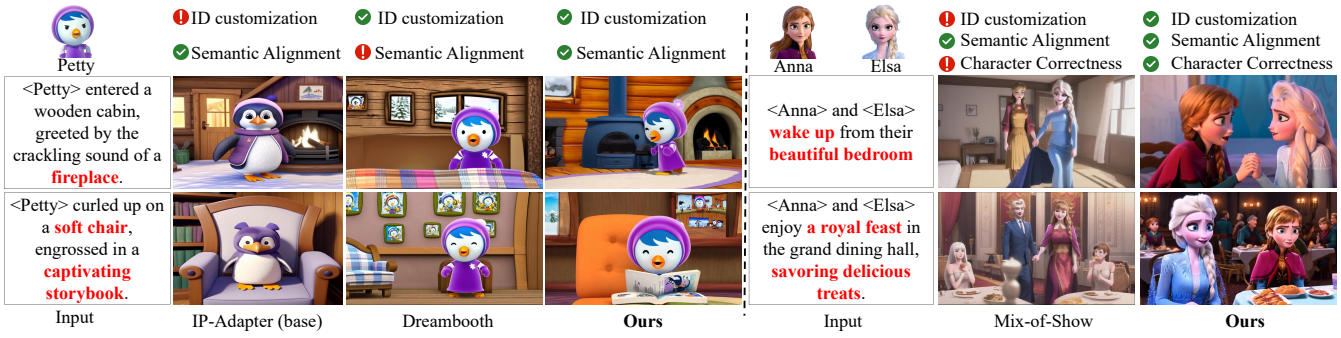
Story Visualization is an emerging task in artificial intelligence with wide-ranging applications in education and entertainment, *e.g.*, comic books creation and movie production (Li et al. 2019; Maharana, Hannan, and Bansal 2022; Zhou et al. 2024b; Cheng et al. 2024). Given a textual narrative and portrait images of characters, the task of story visualization is to generate a series of images visually represent the story. Therefore, the main obstacle in this task is to customize the given characters faithfully and synthesize semantically diverse images that align well with the prompt along the whole storylines (Gong et al. 2023; Su et al. 2023; Song et al. 2020; Chen et al. 2022; Cheng et al. 2024).

Recent diffusion methods for character-consistent image generation (Gal et al. 2022; Ruiz et al. 2023; Ye et al. 2023; Li, Li, and Hoi 2024; Kumari et al. 2023; Han et al. 2023) can be broadly categorized into two types, *i.e.*, adapter-based and customization-based. Adapter-based methods (Ye et al. 2023; Wang et al. 2024; Liu et al. 2024) introduce image-conditioned side networks within diffusion models to provide visual guidance. For example, IP-Adapter (Ye et al. 2023) deploys an image adapter to extract features from input images, while StoryGEN (Liu et al. 2024) incorporates a context module to provide visual context from previous frames. However, these methods struggle with detailed identity extraction and precise character customization. As depicted in Fig. 1(a), IP-Adapter captures the coarse semantics of *Petty*, *e.g.*, a penguin wearing purple, but overlooks finer details of the appearance.

On the other hand, customization-based methods (Ruiz et al. 2023; Hu et al. 2021; Gal et al. 2022; Han et al. 2023) conceptualize the character on a set of customized images, achieving better identity preservation. However, since they are trained in an entangled way, *i.e.*, rely on few text tokens to capture coarse-grained semantics, these methods are overfit and hard to respond to text instructions, also illustrated in Fig. 1(a). Furthermore, their per-concept optimization necessitates a distinct model for each character, resulting in a significant demand for storage resources.

Unlike previous works, we are keen to a unified framework with fine-grained modeling for comprehensive multi-character customization to achieve high-fidelity identity preservation and precise text-semantic alignments. Inspired by ERNIE (Zhang et al. 2019), which enhances language representation through external knowledge, we argue that finer-grained details within each story can be effectively captured with enhanced semantic-rich knowledge. Then we propose a novel Character-Graph (**CG**) to encode fine-grained semantics about the story world, including the given characters, their detailed attributes, and their relations. Characters are presented as object nodes in **CG** with multiple attribute node attached, and their relations serve as the edge to connect all objects. These components collectively define the essence of each story scene. Then each visual story scene can be detailed in text captions through **CG**. By Customiz-

*These authors contributed equally.
†Corresponding author.



(a). Visual comparison for story visualization on single- and multi-character story visualization.

(b). A complete visual story generated by our StoryWeaver.

Figure 1: Our StoryWeaver can achieve high-quality story visualization based on the given characters within a unified model.

ing via Character-Graph enhanced scene-caption pairs (**C-CG**), our proposed model, namely StoryWeaver, is capable of capturing crucial semantics from the story scene, thereby dramatically improving the consistency and alignment on both identity and semantics.

However, multi-character story visualization is still intractable, which suffers from identity blending. This problem stems from an incorrect attention distribution in the diffusion model, where specific character knowledge impacts unrelated regions without spatial constraints. Existing methods (Gu et al. 2024; Hu et al. 2021) adopt a regionally controllable sampling method to address this problem. As these methods necessitate additional spatial inputs, *e.g.*, keypose image and layout, to strictly determine region assignment for different characters, they often encounter conflicts in layout and degradation in identity representation. As shown in the right side of Fig. 1(a), identity variations among the characters "*Anna*" and "*Elsa*" are evident across frames, and the model fails to generate accurate semantics of "*savor treats*".

To remedy this issue, we propose Knowledge-Enhanced Spatial Guidance (**KE-SG**) as external knowledge within the attention mechanism for precise multi-character customization without compromising quality. Specifically, we introduce a knowledge encoder to extract features of different characters, then refine the initial position prior based on the extracted character knowledge to modify the incorrect cross-attention map. Through **KE-SG**, character knowledge from **CG** can accurately attend to the corresponding region in the story scene, ensuring precise identity representation and well-matched text semantics for multi-character visual storytelling tasks. As shown in Fig. 1(b), StoryWeaver achieves vivid visual story plot generation that encompasses both single- and multi-character interactions.

Moreover, we introduce a new benchmark termed TBC-Bench to train our StoryWeaver for both single- and multi-character story visualization, and compare it with a set of state-of-the-art (SOTA) methods (Ye et al. 2023; Liu et al. 2024; Ruiz et al. 2023; Hu et al. 2021; Gu et al. 2024; Yang et al. 2024). The experimental results demonstrate that our StoryWeaver excels not only in character identity preservation, *e.g.,* achieving an average increase of +9.03% DINO-I on single-character visual storytelling, but also in visual quality across various tasks with an average increase of +18.45% CLIP-T and + 19.11% Character F1 for multi-character story visualization.

To sum up, the contributions of this work are three-fold:

- We propose a novel Character Graph to structurally represent semantic-rich knowledge within each story world and a novel StoryWeaver enhanced by the structured knowledge to achieve high-quality visual storytelling.
- We introduce a novel knowledge-enhanced spatial guidance (**KE-SG**) for precise cross-attention assignment to address identity blending, which improves the performance of multi-character generation.
- Our StoryWeaver outperforms a set of compared methods on the newly proposed TBC-Bench in terms of character identity preservation, correct complex scene generation and text semantics alignment.

Related Work

Story Visualization

The task of story visualization is to generate image series aligning with multi-sentence paragraphs while maintaining global semantic consistency throughout the storyline. Earlier works have adapted GAN for this task (Li et al. 2019; Song et al. 2020; Li, Torr, and Lukasiewicz 2022; Li, Kong, and Zhou 2020; Szűcs and Al-Shouha 2022), followed by Transformers-based methods (Chen et al. 2022; Maharana, Hannan, and Bansal 2022) which leverage the long-range dependence properties to enhance semantic coherence.

Recent studies (Feng et al. 2023; Rahman et al. 2023; Su et al. 2023; Liu et al. 2024; Gong et al. 2023; Zhou et al. 2024b; Cheng et al. 2024) have explored diffusion model (Rombach et al. 2022) to achieve consistent image generation, especially for open-ended visual storytelling task(Zhou et al. 2024b; Liu et al. 2024; Cheng et al. 2024). For example, StoryGEN (Liu et al. 2024) trains a visual language context module to extract information from previous-turn images while Storydiffusion (Zhou et al. 2024b) proposes Consistent Self-Attention within a batch.

Image Customization

Image Customization aims to synthesize specific subjects align with given textual contexts. Single-concept customization methods either learn new "word" embeddings from small image sets of customized subjects (Gal et al. 2022; Voynov et al. 2023; Kumari et al. 2023; Dong, Wei, and Lin 2022) or train diffusion models with additional modules to encode visual guidance (Ruiz et al. 2023; Wei et al. 2023; Chen et al. 2024; Jia et al. 2023; Li, Li, and Hoi 2024; Ma et al. 2023; Yuan et al. 2023; Hu et al. 2021).

While significant progress has been made, multi-concept customization remains challenging (Han et al. 2023; Yang et al. 2024; Xiao et al. 2023). To address multi-subject identity blending issue, existing methods use additional loss on large multi-subject dataset or rely on extra optimization efforts to merge multiple models (Xiao et al. 2023; Shentu, Watson, and Moubayed 2024; Zhang et al. 2024; Gu et al. 2024; Xie et al. 2023; Yang et al. 2024). Some methods (Gu et al. 2024; Xie et al. 2023; Zhou et al. 2024a) use spatial control inputs like layout and bounding box annotation to separate characters, but face challenges in natural interaction synthesis and per-character identity preservation.

Method

Overview

In this paper, we aim to achieve fine-grained story world simulation by a unified model for story visualization, as depicted in Fig. 1(b). The overview of Storyweaver is shown in Fig. 2. We first propose the Character-Graph, a novel representation of the semantic-rich knowledge within this particular story world, to enhance consistency generation in diffusion model (Fig. 2(a)). Subsequently, we employ spatial guidance to effectively incorporate the Character-Graph knowledge for precise multi-character generation (Fig. 2(b). In the following sections, we will introduce customization via the Character-Graph (**C-CG**) and then detail the knowledge-enhanced spatial guidance (**KE-SG**).

Customization via the Character-Graph

Existing methods (Ruiz et al. 2023; Hu et al. 2021; Gong et al. 2023; Rahman et al. 2023) use simple tokens for character customization across a few samples to achieve consistent image generation. Inspired by ERNIE (Zhang et al. 2019), which employs an external knowledge map to enhance token representation in language model(**LM**), we integrate a novel Character-Graph to enhance the representation of story scenes, thereby improving character identity preservation and semantic modeling.

Character-Graph Construction. Detailed semantics, including objects, their attributes, and relationships, are crucial to the understanding of visual scenes (Johnson et al. 2015). In the story world, characters act as pivotal **objects** that form the most important part of the world. **Attributes** linked to characters are also crucial, as they vividly depict each character's appearance. Interactions among characters unveil the most intricate dynamics within the story world, representing the unfolding **events** that propel the story forward. As illustrated in Fig. 2(a), Character-Graph in a story world can be formulated as $G =< O, E, A >$, where O represents the character sets, E denotes the set of events, and A refers to the set of attribute nodes associated with O and E.

To construct G, we begin by creating a character vocabulary for the given character set. We collect frontal image I_i for each character, as depicted in Fig. 2(a). Then, we use a vision-language model (**VLM**) to extract a detailed caption with rich semantic from the image, formulated as:

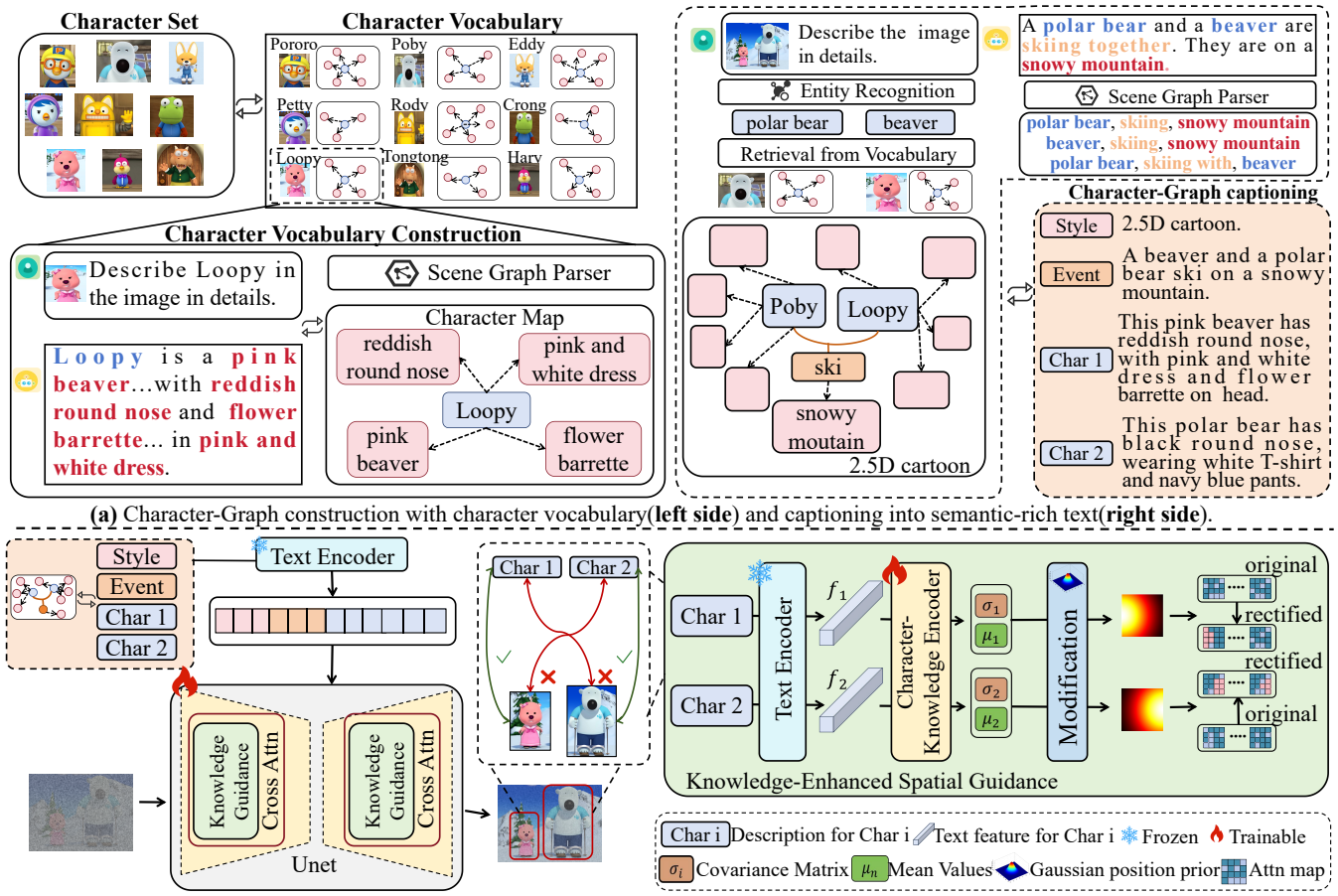
$$C_i = V_{cap}(Instruct_c, I_i), i \in [1, N_c], \tag{1}$$

where C_i is the caption of I_i obtained by prompting the VLM model denoted as V_{cap} with an instruction $Instruct_c$, and N_c is the number of characters in the character set.

However, while such a detailed caption contains rich attributes of the character, it also contains unrelated semantics that are useless for customization. Therefore, we further propose a parsing method to extract detailed semantics related only to the character O_i by:

$$\sum Map < O_i, A_i^k >= (SG_i^{(A)}|O_i) = (Parser(C_i)|O_i). \tag{2}$$

Here $SG_i^{(A)}$ is the scene graph for image I_i obtained by a Scene Graph Parser (Wu et al. 2019) based on C_i, denoted



(a) Character-Graph construction with character vocabulary(**left side**) and captioning into semantic-rich text(**right side**).

(b) Overview of StoryWeaver trained on knowledge-enhanced captions(**left side**) with knowledge-enhanced spatial guidance (**right side**).

Figure 2: Overview. **(a).** We propose Character-Graph to represent semantic-rich knowledge within the story world. **(b).** We enhanced the StoryWeaver with proposed spatial guidance for further improving the performance of mult-character generations.

as $Parser. < O_i, A_i^k >$ is the character map constructed by the character O_i itself with the related k_{th} attribute denoted as A_i^k. As depicted in Fig. 2(a), *"Loopy"* is described by detailed attributes, *e.g., "reddish round nose"*.

Besides, **events**, *e.g.*, "hugging" and "kissing", are used to illustrate the connections among or within characters. Similarly, for a given story scene $\mathcal{F}$ containing character i and character j, we first extract the rich semantic caption $\mathcal{F}_c$ via a **VLM** model, and then a Scene Graph Parser is adopted to extract the event-related semantic:

$$R(O_i, O_j) = (SG_i^{(R)} | (O_i, O_j)) = (Parser(\mathcal{F}_c))_{i,j}, \quad (3)$$

where $\mathcal{F}_c$ is the original scene caption, and $R(O_i, O_j)$ denotes the events between O_i and O_j[1]. Then, Character-Graph that describes the extensive character and event knowledge within the story's scenes can be formulated as:

$$G(O, E) = \{\sum_{i,k} Map < O_i, A_i^k >, \sum_{i,j} R(O_i, O_j)\}. \quad (4)$$

Scene Caption via Character-Graph. Based on Character-Graph, we create a detailed description for each story scene, encompassing the fine-grained attributes of characters and the interactions between them. With this structured and semantic-rich knowledge-enhanced captions, the model can achieve better consistencies of characters and semantic alignment of events.

For each story scene $\mathcal{F}$, we first employ V_{cap} to generate a detailed description prompted by an instruction $Instruct_e$, *e.g.*, "A polar bear and a beaver are skiing on a snowy mountain." Subsequently, we extract the characters and their relations by Scene Graph Parser:

$$\sum_j R_{j,*}, char_j = Parser(V_{cap}(Instruct_e, \mathcal{F})), \quad (5)$$

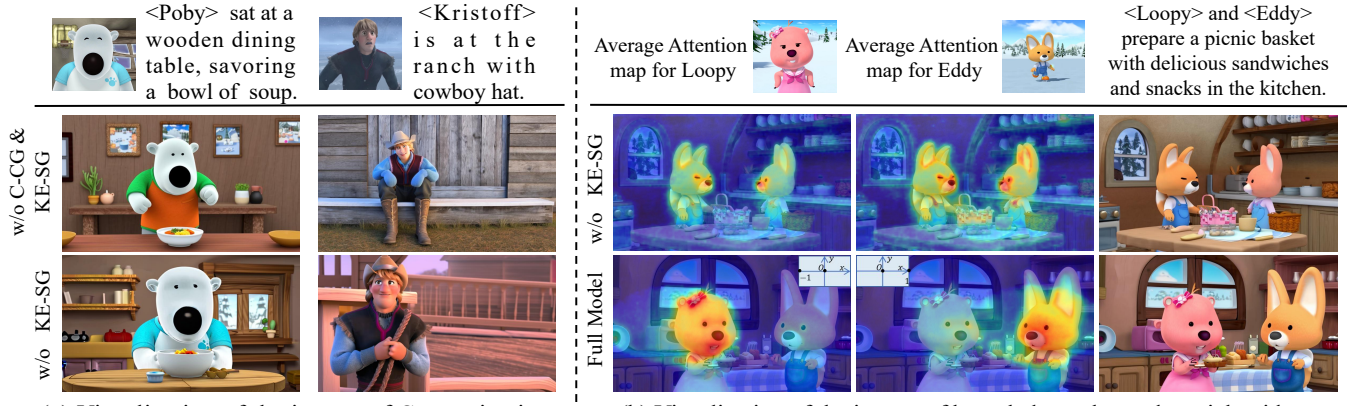
where $char_j$ is the j_{th} character coarse label, *e.g.*, "bear" and "beaver" for the given example, and $R_{j,*}$ refers to the relationship between $char_j$ and object O_*.

Then we lookup the character vocabulary in G to identify the exact characters with the linked attributes formulated as:

$$O_{char}^j = \arg\max_{O_i} Sim(char_j, O_i),$$
$$A_{char}^j = O_{char}^j \otimes \sum Map < O_i, A_i^k > . \quad (6)$$

where O_{char}^j is the matched character of $char_j$ by a match-

[1] O_i and O_j may refer to the same character



(a).Visualization of the impact of Customization via Character-Graph **(C-CG)** .

(b).Visualization of the impact of knowledge-enhanced spatial guidance **(KE-SG)** on attention map.

Figure 3: Visual examples for the impact of Customization via Character-Graph (**C-CG**) and Knowledge-Enhanced Spatial Guidance (**KE-SG**). **(a).**Without **C-CG**, the generator struggles to capture finer-grained details of character. **(b).**Without **KE-SG**, the generator tends to allocate attention uniformly across all regions, resulting in identity blending.

ing function Sim, A^j_{char} denotes the linked attributes associated with O^j_{char} and $\otimes$ refers to the lookup operation.

We then serialize the structured knowledge graph into a scene caption for customization. First, we combine O^j_{char} with the related attributes to generate an appearance description W^j_c for $char_j$ by:

$$W^j_c = O^j_{char} \oplus A^j_{char}, \tag{7}$$

where $\oplus$ denotes the union operation. Next, we use the relationships between objects to describe the events in the scene by W_e, where:

$$W_e = \sum_{j,j'} O^j_{char} \oplus R_{j,j'} \oplus O^{j'}_{char}. \tag{8}$$

Furthermore, we employ a descriptive sentence to characterize the style of all scenes within the story's world, denoted as W_s. Then, the complete story scene caption can be formulated as $T_g = [W_s, W_e, \sum_j^N W^j_c]$ where N is the number of character(s) appear in that scene.

In this case, given scene image frame $\mathcal{F}$ and the obtained T_g, the objective of the diffusion model is:

$$\mathbb{E}_{f\sim E(\mathcal{F}),T_g,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \epsilon_\theta(f_t, t, \tau(T_g))\|^2_2\Big], \tag{9}$$

where ϵ is a random noise, t is the sampled timestep, E is the VAE encoder and τ is the text encoder. We name the model enhanced by C-CG as StoryWeaver because it seamlessly "weave" all elements within G in a unified model and achieve improved customization for story visualization.

Knowledge-Enhanced Spatial Guidance for Multi-Character Generation

Akin to previous studies (Xiao et al. 2023; Yang et al. 2024; Han et al. 2023), StoryWeaver faces challenges with identity blending in multi-character generation.

Attention Blending. In our case, the cross-attention mechanism in the diffusion model that used to update image feature $f_\mathcal{F}$ can be formulated as:

$$Attn = \mathcal{M} \cdot V = \text{Softmax}\Big(\frac{(W_q f_\mathcal{F})(W_k f_T)^T}{\sqrt{d}}\Big) \cdot (W_v f_T), \tag{10}$$

where $\mathcal{M}$ denotes the attention map between text feature f_T and image feature $f_\mathcal{F}$. And W_q, W_k and W_v are learnable projection matrices, and d is the latent projection dimension.

Given $T_g = [W_s, W_e, \sum_j^N W^j_c] = w[1:T]$, the cell $\mathcal{M}^i_{x,y}$ represents the correlations between the word $w[i]$ and the image at pixel (x, y). Considering a situation where $w[i] \in W^j_c$ is part of the description of O^j_{char} with extremely high $\mathcal{M}^i_{x,y}$, where (x, y) should correspond to another character $O^{j'}_{char}$, the incorrect knowledge guidance of $\mathcal{M}^i_{x,y}$ can be highly detrimental to generation of consistent characters. This is evidenced by the visualization in Fig.3(b), where the average attention maps for Eddy's description contribute equally to both regions, leading to identity blending with duplicate character generation.

Knowledge-Enhanced Spatial Guidance. Since identity blending arises from an imperfect attention map, we propose a knowledge-enhanced spatial guidance (KE-SG) to modify the attention maps for precise multi-character generation.

Specifically, we first assign a position prior for character O^j_{char} as an external knowledge, which follows the Gaussian Distribution and can be formulated as:

$$p_j(x, y) = \frac{1}{2\pi\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu_x, y-\mu_y)\Sigma^{-1}(x-\mu_x, y-\mu_y)^T}, \tag{11}$$

where μ_x and μ_y represent the mean values in the horizontal and vertical directions, and Σ denotes the covariance matrix. $p_j(x, y)$ denotes the probability of pixel (x, y) belongs to O^j_{char}. In practice, we assume that characters appearing in the same scene are evenly distributed horizontally. Therefore, we initially set $\mu_y = 0$ and $\Sigma = I$ for all characters while $\mu_x = -1 + j \times \frac{2}{n-1}$ for character O^j_{char}, where n is

Task	Type	Method	# Params/DB(↓)	Pororo			Frozen		
				DINO-I(↑)	CLIP-I(↑)	CLIP-T(↑)	DINO-I(↑)	CLIP-I(↑)	CLIP-T(↑)
Sin-Char	Adapter-based	StoryGEN	1064 M	52.92	76.03	26.98	46.67	72.61	28.05
		IP-Adapter(base)	1038 M	48.85	76.66	29.98	44.15	78.42	31.69
		IP-Adapter(plus)	1063 M	64.36	81.64	24.88	60.87	84.52	27.15
	Customization-based	LORA	1024 M	54.13	75.19	28.53	49.02	82.77	29.18
		Dreambooth	7118 M	61.85	78.86	26.74	55.01	81.07	27.12
		StoryWeaver(ours)	1017 M	**64.96**	**82.65**	**33.26**	**62.17**	**85.24**	**36.74**

Task	Type	Method	# Params/DB(↓)	Pororo			Frozen		
				CLIP-T(↑)	F-Acc(↑)	C-F1(↑)	CLIP-T(↑)	F-Acc(↑)	C-F1(↑)
Multi-Char	Adapter-based	StoryGEN	1064 M	27.27	19.55	27.17	28.91	12.31	21.79
	Customization-based	Mix-of-Show	1164 M	27.20	30.23	44.03	30.71	18.90	30.62
		LoRA-Composer	1425 M	27.86	27.04	47.36	28.88	27.69	39.72
		StoryWeaver(ours)	1017 M	**34.30**	**40.45**	**59.72**	**34.94**	**34.51**	**44.53**

Table 1: Quantitative comparisons on the single- and multi-character generation with existing methods. Our StoryWeaver obviously merits in semantic alignments with high identity customization compared to existing methods.

the total number of characters in the scene[2].

As characters vary in size, we leverage the knowledge from W_c^j to precisely modify the $p_j(x, y)$. Given the corresponding appearance description W_c^j, we propose a knowledge encoder $\mathcal{E}$ to extract the spatial semantics of the character, formulated as:

$$\mathcal{E}(W_c^j) \rightarrow \{(\Delta\mu_x, \Delta\mu_y), \gamma\}, \tag{12}$$

where $(\Delta\mu_x, \Delta\mu_y)$ represent the offset of mean and γ represents the scale of corvariance. Then, the knowledge-enhanced position prior can be obtained by:

$$p_j(x, y|t_j) = \frac{1}{2\pi\sqrt{|\hat{\Sigma}|}} e^{-\frac{1}{2}(x-\hat{\mu_x}, y-\hat{\mu_y})\hat{\Sigma}^{-1}(x-\hat{\mu_x}, y-\hat{\mu_y})^T},$$
$$\hat{\Sigma} = \gamma \cdot \Sigma, \hat{\mu_x} = \mu_x + \Delta\mu_x, \hat{\mu_y} = \mu_y + \Delta\mu_y, \tag{13}$$

where t_j denotes the text feature for W_c^j.

We sample the character-aware spatial guidance $\mathcal{P}^j$ for O_{char}^j by Eq.13. Then, the character-related region on $\mathcal{M}^i$ is enhanced along with the unrelated region decreased by:

$$[\mathcal{M}_t^i]_{(x,y)} = \begin{cases} [\mathcal{M}_t^i]_{(x,y)} + s, \text{if } [\mathcal{P}^j]_{(x,y)} \geq \beta, \\ [\mathcal{M}_t^i]_{(x,y)} - s, \text{if } [\mathcal{P}^j]_{(x,y)} < \beta. \end{cases} \tag{14}$$

where $[\mathcal{M}_t^i]$ refers to the cross-attention map for token $w[i] \in W_c^j$ at t timestep and (x, y) is the pixel coordinate, β is the threshold determine whether image pixel (x, y) is related with O_{char}^j. The time-aware guidance scale s is:

$$s = s(t) = \alpha \cdot (ln(t + 1) + 1), \tag{15}$$

where α is the guidance strength and t is the timesteps. s applies stronger knowledge guidance in the initial steps during the early noisy stages, gradually tapering off the guidance strength to prevent quality degradation.

As shown in Fig.3(b), StoryWeaver achieve correct multi-character generation well-matched the given text instruction.

[2]The coordinate of the position prior is illustrated in Fig.3(b)

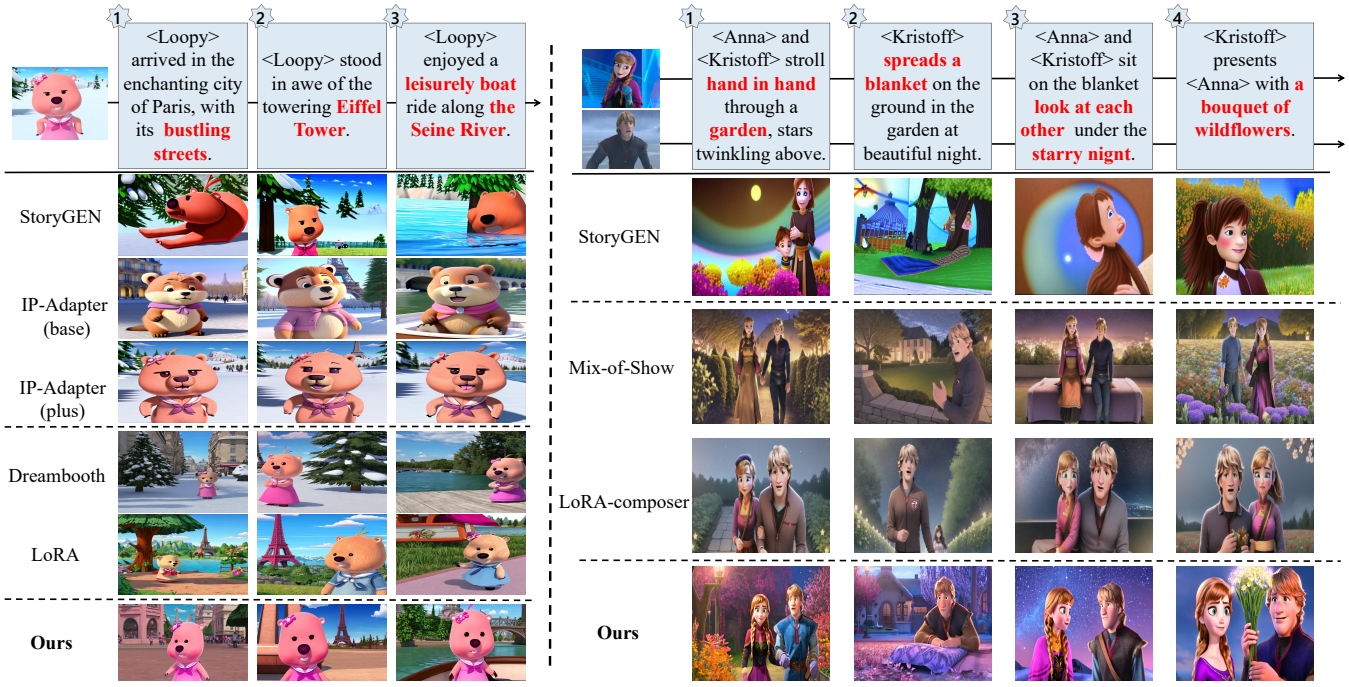
Experiments

Experimental Settings

Dataset Construction Existing datasets (Li et al. 2019; Gupta et al. 2018) for story visualization suffers from low resolution and simple caption annotations. We focus on two popular cartoon series, *i.e.*, *Pororo the Little Penguin* and *Frozen* and construct a dataset featuring multiple cartoon characters. Due to resource constraints, we selected around 10 images per character depicting various events, including over 30 images with multiple characters to capture complex interactions. We also established an evaluation benchmark termed **TBC-Bench**, including 5 single-character stories for each character and 10 multi-character stories in all.

Implement Details We employed Stable-Diffusion v1-5 and implement character knowledge encoder $\mathcal{E}$ as MLP. The whole model is trained with a learning rate of 7×10^{-6} and a batch size of 4. For characters from *Pororo*, we set $\alpha = 2.5$, whereas $\alpha = 1$ for *Frozen* and $\beta = 0.85 \times \|\mathcal{P}^j\|_{max}$. During inference, we employed the LMSDiscrete Scheduler with 100 sampling steps and the text guidance scale is 5.0.

Evaluation Metric Following (Zhou et al. 2024b; Cheng et al. 2024; Gu et al. 2024), we evaluate the methods from three aspects: (1) *Identity Preservation*, *i.e.*, calculating the similarity between generated image and GT character image by DINO (Oquab et al. 2023) (**DINO-I**) and CLIP (Radford et al. 2021) (**CLIP-I**), (2)*Semantic alignment*, *i.e.*, calculating the text-image similarity of the generated image and the text description by CLIP (**CLIP-T**) , (3)*Character Integrity*, *i.e.*, Character F1 (**C-F1**) that measuring the percentage of characters in a generated image that exactly match the story input, and Frame Accuracy (**F-Acc**) that measures whether all characters in a story are present in the generated image. We use a pre-trained DINO as the classifier, where a similarity score exceeding 0.5 between the renderings and the ground truth character image indicates a correct sample.



(a). Visualization on single-character visual storytelling. **(b). Visualization on multi-character visual storytelling.**

Figure 4: The visual comparisons of different methods on single and multi-character visual storytelling. Our StoryWeaver excels in character identity customization and well-matched semantic alignment.

Single Character Customization

Quantitative Results From Tab.1, it is evident that IP-Adapter-plus exhibits the highest character consistency with the poorest text semantic alignment. As shown in Fig. 4(a), its generations are minor editions of the reference image without text semantic alignments. Secondly, the two other customization-based methods achieve better character consistency than adapter-based ones. However, the declination on CLIP-T indicates that such improvement comes at the cost of semantic alignment. Finally, StoryWeaver surpasses all other methods, significantly improving character consistency by +13.02% on DINO-I and text semantic alignment by +15.94% on CLIP-T for *Frozen*, demonstrating the superiority of our proposed Character Graph modeling.

Visualization Results As shown in Fig.4(a), StoryGEN and IP-Adapter (base) struggle to customize characters with intricate details. The generated images bear minimal resemblance to "*Loopy*". Other customization-based methods like Dreambooth and LoRA face challenges in balancing faithful character customization with semantic alignment. Conversely, with semantic-rich knowledge enhanced, StoryWeaver excels in crafting high-quality images with faithful identity preservation, aligning well with the prompts.

Multi Character Customization

Quantitative Results From Tab.1, we can first find that existing methods all fails to achieve semantic-aligned generation. However, our proposed Storyweaver significantly enhances text semantic alignments (+23.12% on *Pororo* and +13.77% on *Frozen*) while also notably improving character consistency (+26.10% in C-F1 on *Pororo* and + 12.11% on *Frozen*). Experimental results demonstrate that our unified modeling and **KE-SG** substantially enhances the efficacy and stability of multi-character generation.

Visualization Results As shown Fig.4, StoryGEN struggles with responding the given prompt, mainly due to the absence of explicit learning for the character and event. For instance, neither "*Anna*" nor "*Kristoff*" are generated correctly in its renderings. Multi-concept tuning methods are much better in identity preservation. However, due to the strong input spatial constraints of keypose or layout, these two methods synthesize unnatural interaction and barely achieve aligned text semantics, as evidenced by the depictions of "*starry night* " and "*wildflowers*". Conversely, StoryWeaver can well respond to the text semantics while retaining high character consistency.

Conclusion

In this paper we introduce StoryWeaver, a unified model with intricate characters customization for story visualization. We first propose a novel Character-Graph that encapsulate semantic-rich knowledge within the story world to enhance our StoryWeaver. Then, we introduce knowledge-enhanced spatial guidance to refine cross-attention maps for precise multi-character generation. Experiment results demonstrates that StoryWeaver achiever better hitherfidelity in identity customization and better semantic alignment than a set of single- and multi-customization methods.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. U21B2037, No. 62072389, No. 62302411), the Natural Science Foundation of Fujian Province of China (No.2021J06003), China Postdoctoral Science Foundation (No. 2023M732948), and partially sponsored by CCF-NetEase ThunderFire Innovation Research Funding (NO. CCF-Netease 202301).

References

Chen, H.; Han, R.; Wu, T.-L.; Nakayama, H.; and Peng, N. 2022. Character-centric story visualization via visual planning and token alignment. *arXiv preprint arXiv:2210.08465*.

Chen, W.; Hu, H.; Li, Y.; Ruiz, N.; Jia, X.; Chang, M.-W.; and Cohen, W. W. 2024. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36.

Cheng, J.; Lu, X.; Li, H.; Zai, K. L.; Yin, B.; Cheng, Y.; Yan, Y.; and Liang, X. 2024. AutoStudio: Crafting Consistent Subjects in Multi-turn Interactive Image Generation. *arXiv preprint arXiv:2406.01388*.

Dong, Z.; Wei, P.; and Lin, L. 2022. Dreamartist: Towards controllable one-shot text-to-image generation via contrastive prompt-tuning. *arXiv preprint arXiv:2211.11337*.

Feng, Z.; Ren, Y.; Yu, X.; Feng, X.; Tang, D.; Shi, S.; and Qin, B. 2023. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; He, Y.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; et al. 2023. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Gupta, T.; Schwenk, D.; Farhadi, A.; Hoiem, D.; and Kembhavi, A. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, 598–613.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7323–7334.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jia, X.; Zhao, Y.; Chan, K. C.; Li, Y.; Zhang, H.; Gong, B.; Hou, T.; Wang, H.; and Su, Y.-C. 2023. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. *arXiv preprint arXiv:2304.02642*.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, B.; Torr, P. H.; and Lukasiewicz, T. 2022. Clustering generative adversarial networks for story visualization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 769–778.

Li, C.; Kong, L.; and Zhou, Z. 2020. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, 73: 102956.

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

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6329–6338.

Liu, C.; Wu, H.; Zhong, Y.; Zhang, X.; Wang, Y.; and Xie, W. 2024. Intelligent Grimm-Open-ended Visual Storytelling via Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6190–6200.

Ma, Y.; Yang, H.; Wang, W.; Fu, J.; and Liu, J. 2023. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*.

Maharana, A.; Hannan, D.; and Bansal, M. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, 70–87. Springer.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rahman, T.; Lee, H.-Y.; Ren, J.; Tulyakov, S.; Mahajan, S.; and Sigal, L. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2493–2502.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 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*, 22500–22510.

Shentu, J.; Watson, M.; and Moubayed, N. A. 2024. AttenCraft: Attention-guided Disentanglement of Multiple Concepts for Text-to-Image Customization. *arXiv preprint arXiv:2405.17965*.

Song, Y.-Z.; Rui Tam, Z.; Chen, H.-J.; Lu, H.-H.; and Shuai, H.-H. 2020. Character-preserving coherent story visualization. In *European Conference on Computer Vision*, 18–33. Springer.

Su, S.; Guo, L.; Gao, L.; Shen, H. T.; and Song, J. 2023. Make-A-Storyboard: A General Framework for Storyboard with Disentangled and Merged Control. *arXiv preprint arXiv:2312.07549*.

Szűcs, G.; and Al-Shouha, M. 2022. Modular StoryGAN with background and theme awareness for story visualization. In *International Conference on Pattern Recognition and Artificial Intelligence*, 275–286. Springer.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. p+: Extended textual conditioning in text-to-image generation. *arXiv preprint arXiv:2303.09522*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 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*, 15943–15953.

Wu, H.; Mao, J.; Zhang, Y.; Jiang, Y.; Li, L.; Sun, W.; and Ma, W.-Y. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6609–6618.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Yang, Y.; Wang, W.; Peng, L.; Song, C.; Chen, Y.; Li, H.; Yang, X.; Lu, Q.; Cai, D.; Wu, B.; et al. 2024. LoRA-Composer: Leveraging Low-Rank Adaptation for Multi-Concept Customization in Training-Free Diffusion Models. *arXiv preprint arXiv:2403.11627*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yuan, Z.; Cao, M.; Wang, X.; Qi, Z.; Yuan, C.; and Shan, Y. 2023. Customnet: Zero-shot object customization with variable-viewpoints in text-to-image diffusion models. *arXiv preprint arXiv:2310.19784*.

Zhang, Y.; Yang, M.; Zhou, Q.; and Wang, Z. 2024. Attention Calibration for Disentangled Text-to-Image Personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4764–4774.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced Language Representation with Informative Entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451.

Zhou, D.; Li, Y.; Ma, F.; Zhang, X.; and Yang, Y. 2024a. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6818–6828.

Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024b. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *arXiv preprint arXiv:2405.01434*.