# Personalized Comic Story Generation

**Wenxuan Peng**[1], **Peter Schaldenbrand**[2], **Jean Oh**[2]✉,
[1]Nanyang Technological University, Singapore
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

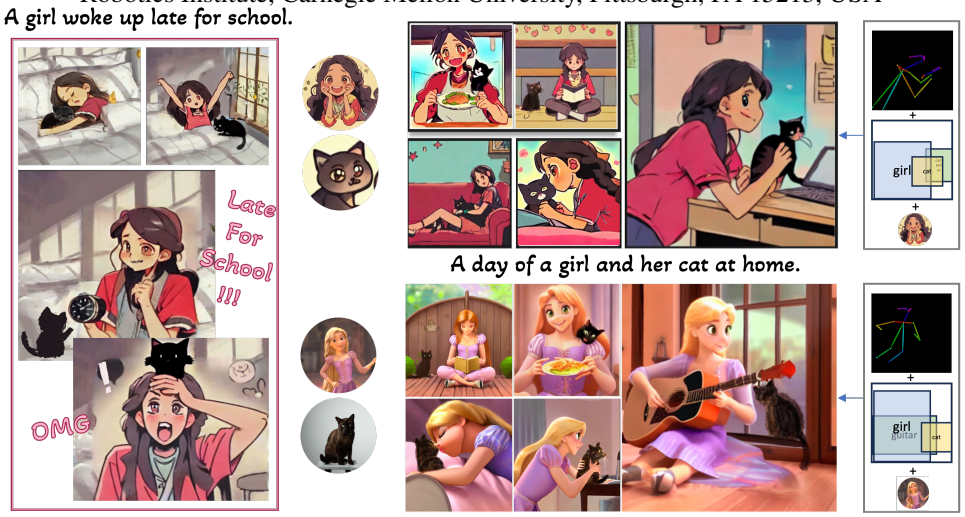


Figure 1: **PCSG Results** Generated comic strip using the proposed model, illustrating its ability to produce diverse, controllable, and personalized comic narratives from user-provided plots.

## Abstract

We introduce PCSG, a diffusion-based text-to-image synthesis framework for supporting comic story generation, a domain in which authors require control over the consistency, composition, and diversity of content. To support these three requirements, PCSG has controllable plugins for (1) character consistency, (2) scene layout specification, and (3) character pose specification. The novel combination of these plugins enables users to exert fine-grained control and manifest their envisioned comic narratives with personalized characters. Our system provides flexibility which greatly improved user satisfaction in our study over existing approaches such as using MidJourney or Stable Diffusion. To further advance this field and facilitate community engagement, we will open source our code soon.

**1. Introduction** Recent advancements in text-to-image synthesis models have vividly demonstrated the ability to convert narrative descriptions into compelling visual stories [1, 2, 3, 4, 5, 6]. Originating with the approach of StoryGAN [1], the emphasis largely remained on maintaining character consistency, sometimes sacrificing diversity and relegating characters to repetitive, standard poses. As the domain evolved, the emergence of diffusion-based methods, exemplified by StoryDALL·E [4] and Make-a-Story [6], notably elevated the visual quality of the synthesized images. While these approaches are adept at visualizing stories on basic cartoon datasets, often fall short in delivering a richly diverse and personalized visual narrative tailored to the user's unique vision. Addressing this gap, we present the Personalized Comic Story Generation (PCSG). Envisioned as a multi-modal collaborative diffusion-based framework, PCSG seamlessly integrates specialized modules—character consistency, layout-to-image, and pose-to-image—into a holistic diffusion pipeline. Beyond replicating character consistency, PCSG places control firmly in the user's hands, enabling them to express their entire narrative. Our goal is to create a cohesive visual story that not only aligns with user-defined aesthetics but also offers significant controllability over the narrative's finer details.

Submitted to NeurIPS 2023 Workshop on Machine Learning for Creativity and Design

## 2. Methodology

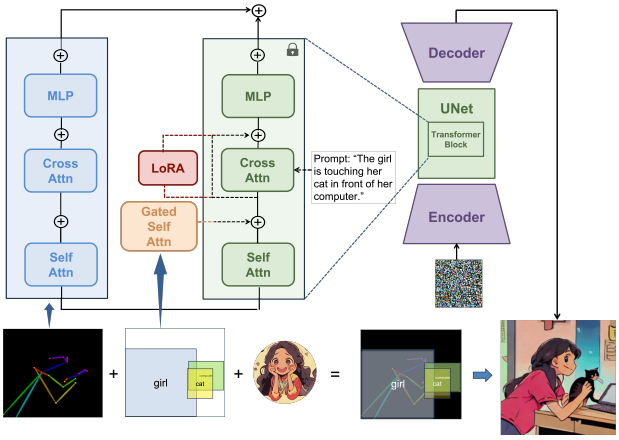


Figure 2: PCSG model architecture

Our PCSG framework represents a non-trivial innovation in the field of text-to-image synthesis, situated atop the robust Stable Diffusion model [7]. The distinctive ingenuity of PCSG lies in its synergistic incorporation of three disparate control modules *LoRA* [8], *GLIGEN* [9], and *ControlNet* [10]—each optimized for separate tasks and originally implemented on distinct codebases.

***a. Character Consistency Plugin*** Initially, by employing Low Rank Adaptation (LoRA), we streamline our model by selectively fine-tuning a small subset of parameters. Formally, given a weight matrix $W \in \mathbb{R}^{m \times n}$, LoRA decomposes it into a product of two low-rank matrices, $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{k \times n}$, where $k \ll \min(m, n)$. LoRA's modifications of low-rank matrices ensures efficient adaptation without information loss. Our adoption of LoRA is not merely plug-and-play; it necessitates meticulous mapping and alignment due to differing initialization structures across codebases. Additionally, inspired by DreamBooth [11], we we use rare-word combinations (eg. comic *xxy5syt00* girl) to create a distinct association between each token and character identity allowing generated comic panels to contain user specified characters.

***b. Layout-To-Image Plugin*** Having embedded our personalized characters within the model, we shift our focus to their precise spatial placement on the canvas, inclusive of background objects. Bounding box layouts created by users specify composition, as they are an intuitive representation. In the pursuit of precision, we employ GLIGEN. Tailored for layout-to-image tasks, GLIGEN integrates a unique gated self-attention mechanism within the Stable Diffusion model. Notably, during training on the layout2image task, only this gated self-attention layer—positioned between the model's original self-attention and cross-attention layers—is learned, ensuring an adept interpretation of the user-defined layouts.

***c. Pose-To-Image Plugin*** However, spatial placement alone doesn't cater to the nuanced character actions and poses, like "holding an item" or "playing guitar". Characters are often squeezed into bounding boxes, compromising visual fidelity. To address this, we integrate the Pose-to-Image plugin backed by the ControlNet—a system adapts the stable diffusion model to interpret additional visual inputs for intricate pose controls. ControlNet trains a copy of the stable diffusion model's parameters on pose2image datasets, all while keeping the core model intact. While primarily designed for single subject pose regulation, we innovatively combined its capabilities with our multi-object layout-to-image system, facilitating intricate pose controls even in complex multi-object scenes.

***3. Experiments and Results*** Further demonstrations of PCSG's versatility across styles can be found in Appendix section 0.2. We conducted a user study to compare PCSG against its base Stable Diffusion model, PCSG without pose control, and a popular text-to-image platform MidJourney, highlighting the significance of user control in the image generation process. More details are presented in Appendix section 0.3

***4. Discussion*** Combining these modules is not straightforward; it requires rigorous alignment, ablation studies, and extensive validations to harmonize their diverse capabilities, synergizing a '1+1>2' effect. Illustrated in Appendix section 0.1, we observe that while layout control improves composition, it could undermine fidelity by squeezing characters to limited areas. Introducing pose control effectively counters this issue, elevating the overall visual quality. This showcases the potential of modular combinations in multi-modal collaborative diffusion models. It's noteworthy that despite PCSG's visual quality being inferior to the industry leaders like Midjourney, user feedback still favored PCSG for its ability to capture their envisioned comic panels with higher fidelity. This underscores a pivotal insight: user participation and control in the image generation process can have a more profound impact on perceived fidelity than the sheer visual quality of the generated images.

***5. Ethical Considerations*** PCSG is influenced by its underlying model, Stable Diffusion, that may contain biases, potentially reinforcing harmful stereotypes. Users should be aware of these biases and refrain from using PCSG for misinformation or harmful content creation.

## References


[1] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019. 1

[2] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*, 2021. 1

[3] Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. Character-centric story visualization via visual planning and token alignment. *arXiv preprint arXiv:2210.08465*, 2022. 1

[4] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, pages 70–87. Springer, 2022. 1

[5] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*, 2022. 1

[6] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2493–2502, 2023. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[9] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[10] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 2

[11] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

## Supplementary material. Additional experiments and results

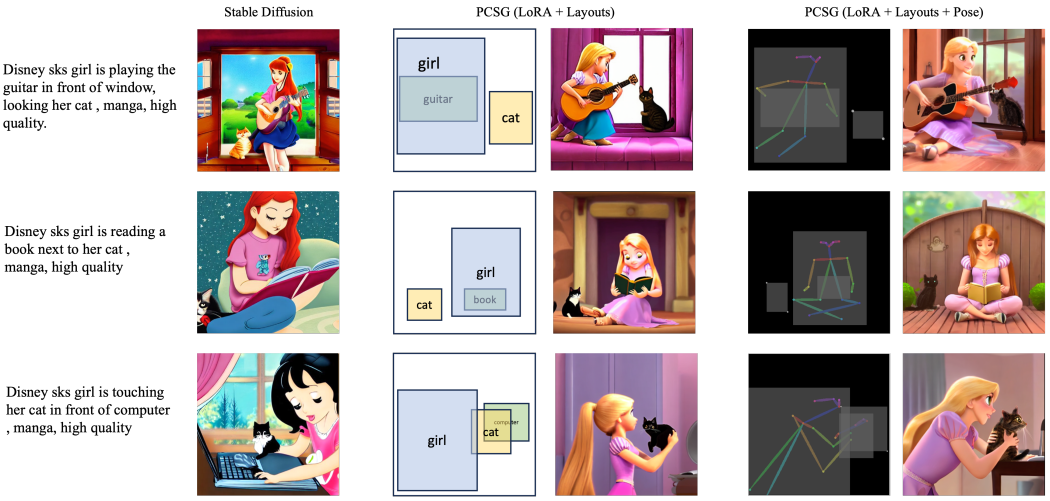


Figure 3: Comparison of generated comic panels using vanilla stable diffusion, PCSG with and without pose control and full PCSG pipeline

### 0.1 Control Ablation Study.

Figure 3 depicts the individual effects of adding our controlability plugins. In the figure, the first column exhibits the baseline performance of the original stable diffusion v1-5 model (the base of our PCSG framework), which is limited in terms of both personalization and structural articulation. The second column improves upon this by incorporating both the “Character Consistency" and “Layout-To-Image" modules, which ameliorate character and layout rendition but still face challenges like the characters being overly constrained and squeezed within their bounding boxes, compromising visual fidelity. In contrast, the third column showcases the output of our complete PCSG pipeline, enriched with the “Pose-To-Image" module. The collective influence of text, spatial, and pose controls yields comics of significantly elevated quality, accurately capturing the characters’ nuanced behaviors and interactions. These experiments underline the incremental benefits brought about by each module, validating the efficacy of PCSG’s comprehensive approach to personalized comic generation.

### 0.2 More Results.

In the next page, we showcase more qualitative results generated from a combination of **text, layout, pose, character**. We also demonstrate that our model has the ability to generalize to different styles.

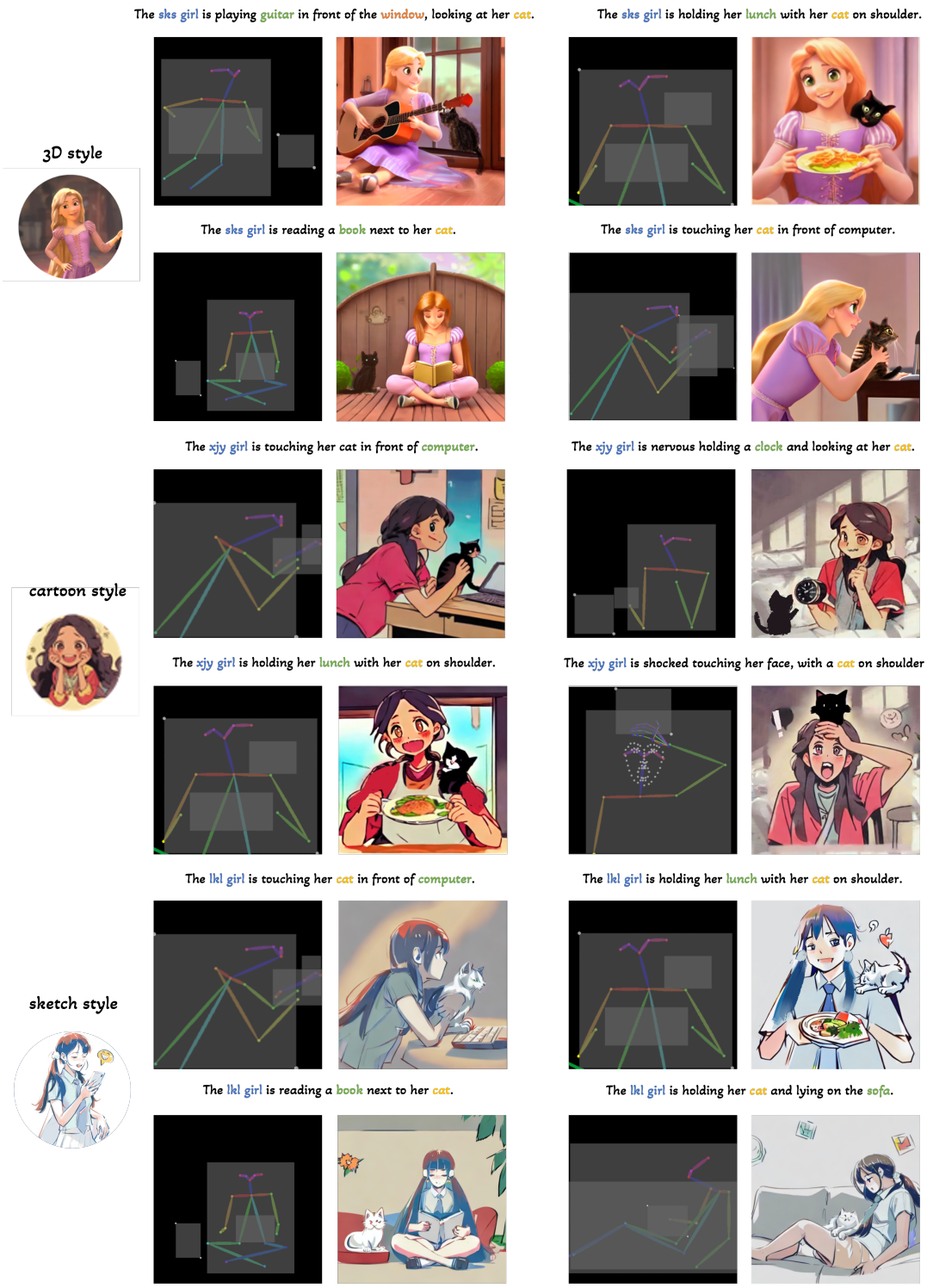


Figure 4: Depiction of multiple generated comic panels, demonstrating the model's capability to produce diverse and personalized character depictions guided by the integrated inputs of scene layout, pose, and narrative context. *sks*, *xjy*, *lkl* are rare tokens that used to embed the personalized characters in the model. Colored words are specified with layout inputs.

### 0.3 User Study

To perceptually evaluate the efficacy of our proposed methods, we conducted a user study employing 30 human evaluators.

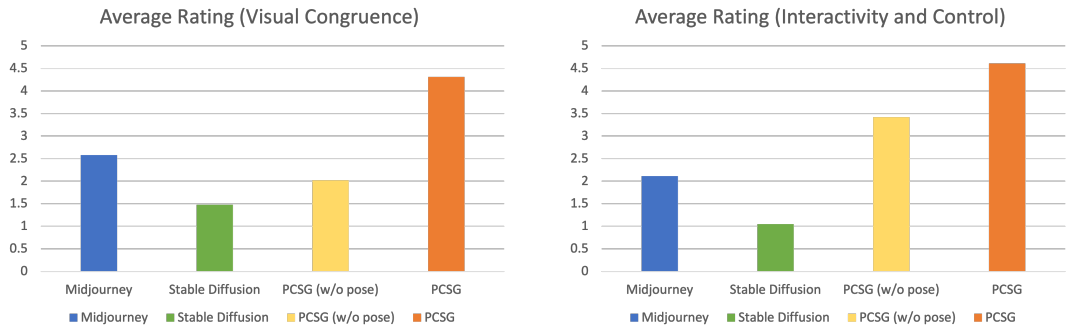


Figure 5: Comparative User Satisfaction Scores across Different Image Generation Platforms.

**Study Design**

Participants were presented with several character reference images and were instructed to craft personalized comic panels. Firstly, participants were asked to envision a particular comic panel and articulate their visions as scripts or prompts. Subsequently, four platforms were tested: Midjourney, Stable Diffusion v1-5, PCSG (w/o pose) (equipped solely with the Character Consistency and Layout2Image plugins), and PCSG (our full pipeline).

**Evaluation Metrics**

Participants rated their satisfaction based on two primary criteria:

1. **Visual Congruence:** The resemblance of the generated image to the participant's envisioned scene.
2. **Interactivity and Control:** How involved participants felt during the generation process, gauging their perceived level of interaction and influence over the platform's outcomes.

Satisfaction was quantified on a scale of 1 to 5, with 1 being least satisfied and 5 being most satisfied.

**Results**

Our findings revealed:

- Midjourney & Stable-Diffusion, although producing aesthetically pleasing images, often fell short in accurately capturing participants' visions, resulting in lower control scores. For example, one user noted, "While Midjourney's outputs are undeniably high-quality and captivating, I faced significant challenges in aligning the character's actions with my instructions. Often, the character's pose or view angle didn't align with my envisioned scene. Also, sometimes, the style felt a tad too 'Midjourney', straying from my provided character reference."
- PCSG (without pose), despite being a trimmed version of our full pipeline, frequently compromised on image fidelity and struggled with accurate pose generation.
- PCSG (full pipeline) emerged as the most preferred platform, with the majority of participants expressing heightened satisfaction with its capability to capture intricate details and provide superior control over the generation process. As one participant enthusiastically shared, "PCSG feels tailor-made for my creative needs! It's as if my visions are instantly transformed into these delightful panels. The idea of using poses, layouts, and text inputs together is very user-friendly and powerful."

## 0.4 Comparative Advantages of PCSG

To provide a more tangible understanding of the user experiences reflected in our user study, we present a comparative visualization of generated results between PCSG and the state-of-the-art platform, *Midjourney*.

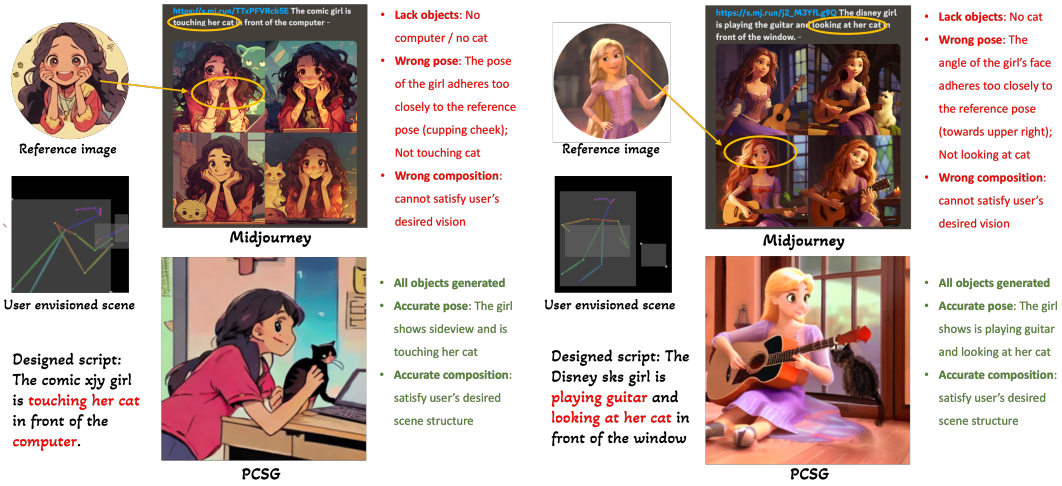


Figure 6: Comparative visualization of *Midjourney* and *PCSG* outputs. Despite *Midjourney*'s commendable image generation quality, it grapples with capturing the user's envisioned pose and composition for this specific task. In contrast, *PCSG* adeptly captures and materializes the user's specific requirements, highlighting its paramount efficiency in personalized comic story generation.

In our analysis with *Midjourney*, while it produces high-quality images, its black-box nature offers users limited control. Outputs often adhere too rigidly to the reference image's pose and can misinterpret user prompts, occasionally overlooking important scene objects. Refinements necessitate restarting the entire generation process, making iterations cumbersome. On the other hand, PCSG is designed to foreground the user's creative intent. With enhanced controllability and precision, it allows users to craft specific character poses and scene compositions, ensuring outputs that are both aesthetically pleasing and aligned with the visual narratives in users' minds. Our user study corroborates this, revealing that participants found PCSG to more effectively capture their envisioned scenes.