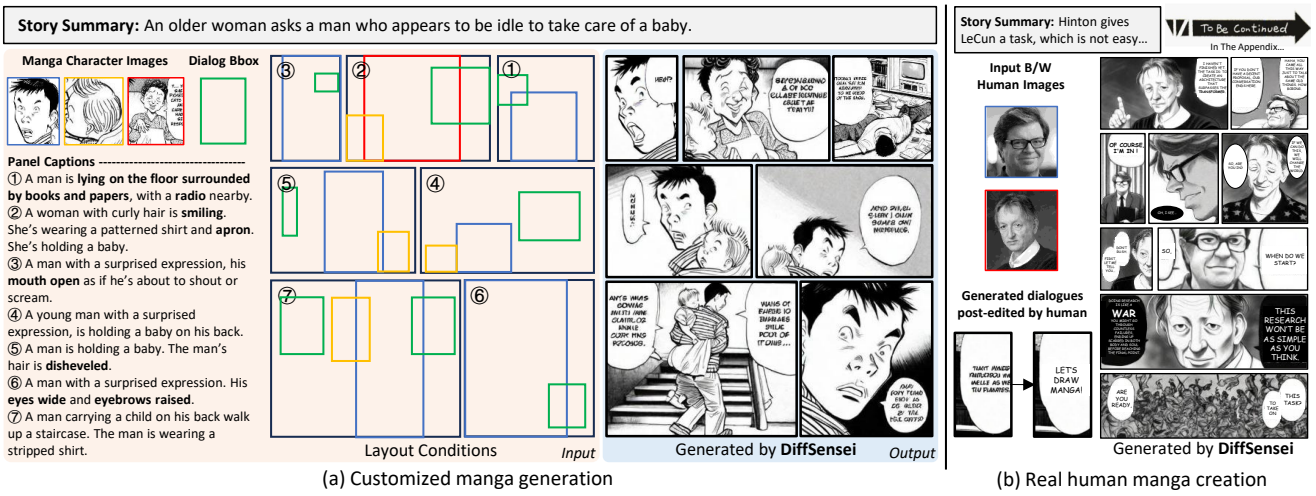
DiffSensei: Bridging Multi-Modal LLMs and Diffusion Models for Customized Manga Generation

Jianzong Wu[1,2] Chao Tang[1] Jingbo Wang[2] Yanhong Zeng[2] Xiangtai Li[3,4] Yunhai Tong[1]
[1] Peking University [2] Shanghai AI Laboratory [3] Nanyang Technological University [4] Bytedance Seed
Project Page: https://jianzongwu.github.io/projects/diffsensei/
Email: jzwu@stu.pku.edu.cn, xiangtai94@gmail.com



(a) Customized manga generation (b) Real human manga creation

Figure 1. Results of **DiffSensei**. (a) Customized manga generation with controllable character images, panel captions, and layout conditions. Our DiffSensei successfully generates detailed character expressions and states following the panel captions. (b) Manga creation for real human images. The dialogues are post-edited by humans. The continuation is in the Appendix. We **strongly** recommend that the readers see the Appendix for more comprehensive results. Manga reading order: Right to left. Top to bottom.

Abstract

*Story visualization, the task of creating visual narratives from textual descriptions, has seen progress with text-to-image generation models. However, these models often lack effective control over character appearances and interactions, particularly in multi-character scenes. To address these limitations, we propose a new task: **customized manga generation** and introduce **DiffSensei**, an innovative framework specifically designed for generating manga with dynamic multi-character control. DiffSensei integrates a diffusion-based image generator with a multimodal large language model (MLLM) that acts as a text-compatible identity adapter. Our approach employs masked cross-attention to seamlessly incorporate character features, enabling precise layout control without direct pixel transfer. Additionally, the MLLM-based adapter adjusts character features to align with panel-specific text cues, allowing flexible adjustments in character expressions, poses, and actions. We also introduce **MangaZero**, a large-scale dataset tailored to this task, containing 43,264 manga pages and 427,147 annotated panels, supporting the visualization of varied character interactions and movements across sequential frames. Extensive experiments demonstrate that DiffSensei outperforms existing models, marking a significant advancement in manga generation by enabling text-adaptable character customization. The code, model, and dataset are open-sourced to the community.* [1]

1. Introduction

Story visualization, the process of generating visual narratives from textual descriptions, is a rapidly evolving field [5, 18, 21, 25, 37, 46, 53]. Among its various applications, manga generation holds particular significance due

[1]The work is done in Shanghai AI Laboratory. Corresponding to: Xiangtai Li, Jingbo Wang.

to its popularity and unique narrative requirements. Unlike traditional story visualization, manga demands consistent characters across panels, precise layout control for positioning multiple characters, and the seamless integration of dialogues in a coherent, visually engaging manner.

Currently, manga generation remains an underexplored area. Most existing research focuses on low-level image-to-image tasks, primarily converting general images to a manga style [20, 33, 48, 50]. While these tasks enhance the visual appeal of static images, they do not extend to generating fully customized manga content from scratch. In general story visualization, current approaches have demonstrated some success in generating coherent image sequences from text. Still, they cannot often customize the characters across scenes [21, 25, 46, 53], a critical requirement for manga generation. Additionally, they do not provide the necessary control over layouts and dialogue placements, which are also essential for manga. A likely reason for these limitations is that existing story visualization datasets typically lack character annotations and layout controls [15, 18, 21, 46]. Another research direction has explored zero-shot character customization, showing promise for achieving character customization across manga panels [4, 9, 19, 29, 36, 39, 47]. However, these approaches often result in rigid "copy-pasting" effects [4, 39, 47], which limit expressive character variation and detract from narrative depth. This limitation largely stems from the scarcity of datasets capturing multiple appearances of the same character in varied expressions and poses.

To address these limitations, we propose a new task: **customized manga generation**. As illustrated in Fig. 1, this task focuses on creating manga images with multiple characters, each customized based on a character image and positioned according to user input. Characters must dynamically adapt to text prompts, altering their expressions, motions, and poses as the narrative unfolds. Dialog layout should also be managed to generate expressive manga panels. Compared to traditional story visualization tasks, this proposed task prioritizes essential manga-specific controllability, aiming to generate vivid manga panels that are both coherent and visually engaging while supporting the customization of multiple characters. To address the lack of a dedicated dataset for customized manga generation, we collect a dataset of Japanese black-and-white manga, forming the basis of the proposed task. The resulting dataset, **MangaZero**, is the first large-scale collection designed to support multi-character, multi-state manga generation.

To tackle this task, we introduce a novel framework, **DiffSensei**, which leverages a diffusion-based image generator to produce customized manga panels. However, we observe that, even with training on multiple appearances of the same character, the generated portraits often tend to rigidly follow the pixel distribution of the input character image, resulting in limited variations in appearance, pose, and motion based on the text input. Inspired by recent advancements in image editing using Multimodal Large Language Models (MLLMs) [6, 8, 16, 34, 40–43, 46, 49], we propose using an MLLM as a text-compatible character adapter. This approach enables seamless, dynamic adjustments to characters in response to textual cues, thereby supporting coherent and expressive manga panel generation. Additionally, we incorporate masked attention injection to manage character layout, along with a dialog embedding technique tailored specifically for manga, allowing for precise control over dialog placement. Through extensive experimentation, we validate DiffSensei's capability to generate coherent, expressive manga panels that maintain narrative consistency and offer improved character control. This represents a significant advancement in story visualization.

- We introduce a new task: **customized manga generation**, focused on generating manga images with multiple characters, each dynamically adapting to text prompts and positioned according to layout specifications.
- We present **MangaZero**, the first large-scale dataset specifically designed for multi-character, multi-state manga generation, addressing a significant gap in story visualization training data. The dataset will be released for the image generation community.
- We propose **DiffSensei**. As far as we know, it is the first framework for customized manga generation that links diffusion models and MLLMs. The MLLM serves as an adaptable character feature adapter, enabling characters to respond dynamically to textual cues. Extensive experiments demonstrate the effectiveness of DiffSensei.

2. Related Work

Story visualization. Story visualization, the process of generating visual narratives based on given stories, is rapidly evolving. Many approaches can generate consistent image sequences derived from story content [5, 21, 25, 37, 46, 53]. Despite recent advancements, the field faces significant limitations. Most existing methods generate story images solely from text and image-level prompts [21, 25, 46, 53], which restricts control over individual characters. This limited control over characters reduces the flexibility and depth of story visualization. A key factor is that current training datasets [10, 15, 18, 21, 46] lack character-specific annotations. In response to the data limit, recent works [5, 37] explore multi-character control using training-free methods that leverage existing subject preservation techniques, such as IP-Adapter [47]. Other works [4, 7, 9, 11, 12, 19, 29, 32, 36, 39, 44, 47] try to train diffusion models for a multi-character customized generation. However, these approaches often result in a "copy-pasting" effect, restricting the diversity of expressions and actions needed for dynamic storytelling. For training-free

Table 1. **Comparison between MangaZero and related publically available datasets.** A story is defined as a sequence of continuous images annotated consistently with character IDs. In MangaZero, a story means a manga page. A panel means a distinct story image, or called frame [21, 46]. Most series in MangaZero are still popular in 2024. Please see the Appendix for the dataset details.

Dataset	Type	Resolution	#Series	#Stories	#Panels	Annotations			Origin
						Caption	Character	Dialog	
PororoSV [18]	Animation	Fix	1	15,336	73,665	✓	×	×	2003-2016
FlintstonesSV [10]	Animation	Fix	1	25,184	122,560	✓	✓	×	1960-1966
StorySalon [21]	Animation	Fix	446	18,255	159,778	✓	×	×	YouTube
StoryStream [46]	Animation	Fix	3	12,614	257,850	✓	×	×	1939-2013
Manga109 [2]	B/W Manga	Vary	109	10,602	103,850	×	✓	✓	1970-2010
MangaZero	B/W Manga	Vary	48	**43,264**	**427,147**	✓	✓	✓	1974-2024

methods, combining multiple models can significantly slow down inference speeds. To address these challenges, we first introduce a large-scale manga generation dataset with finely curated character annotations and then develop a novel framework utilizing diffusion models and MLLMs that enables the dynamic generation of manga panels.

Manga generation. The field of black-and-while manga generation has received limited exploration. Most existing researches focus on low-level image-to-image tasks, primarily transferring general images to a manga style [20, 33, 48, 50]. Recent works contribute to manga content understanding [30, 31, 35]. In contrast, we propose the customized manga generation task beyond style transfer to offer complete character and story-driven manga generation.

MLLMs for personalized image generation. MLLMs have shown remarkable potential in personalized image generation, particularly for tasks involving image editing and customization [3, 6, 8, 16, 34, 40–42, 46, 49]. Notably, CAFE [52] explores customizing subject appearances through textual instructions. However, MLLM-driven image generation for multi-character narratives remains an open challenge, primarily due to the difficulties in maintaining inter-character relationships and scene continuity. Our framework proposes an MLLM-based identity adapter that enhances dynamic story consistency in multi-character manga generation. In contrast to previous works, our framework takes multi-character features as input and edits these features collectively, following the text prompt, enabling flexible subject editing across multiple characters.

3. The MangaZero Dataset

In this section, we first define the problem in Sec. 3.1. Then, we introduce the dataset construction pipeline in Sec. 3.2.

3.1. Problem Formulation

We introduce a challenging new task: customized manga generation. This task focuses on generating manga images where multiple characters, each with their distinct image inputs, are customized and positioned by users. Importantly, characters must adapt to the text prompts by modifying their expressions, motions, and poses dynamically, even when only a limited set of character images is available. To generate a manga story across N panels (or frames), the inputs include: text prompts for each panel $T_0, T_1, \ldots, T_{N-1}$, character images $\mathbf{I} = I_0, I_1, \ldots, I_{K-1}$, character bounding boxes for each panel $\mathbf{B}_0^c, \mathbf{B}_1^c, \ldots, \mathbf{B}_{N-1}^c$, and dialogue bounding boxes for each panel $\mathbf{B}_0^d, \mathbf{B}_1^d, \ldots, \mathbf{B}_{N-1}^d$. The visualization of a panel is represented as $P_i = \Phi_\theta(T_i, \mathbf{I}, \mathbf{B}_i^c, \mathbf{B}_i^d)$, where Φ is the overall model function and θ represents the model's learned parameters.

Discussion. This task diverges from existing story visualization and continuation tasks [21, 25]. Specifically, in story visualization tasks, a panel is produced using $P_i = \Phi_\theta(T_i)$, while in story continuation tasks, the panel generation depends on previous panels as $P_i = \Phi_\theta(T_i, T_{i-1}, P_{i-1})$ for $i > 0$. Both lack explicit character control, a crucial element in storytelling. Furthermore, the proposed task differs from subject-driven image generation approaches [29, 39, 47], as it demands that models not only generate accurate character representations but also modify characters' attributes in response to panel captions and layouts, resulting in varied and coherent narrative visuals. Our experiments, detailed in Sec. 5, demonstrate that our model significantly surpasses baseline models in these critical aspects.

3.2. Dataset Construction

In this section, we introduce the proposed large-scale manga story visualization dataset **MangaZero**.

Comparison with related datasets. A comprehensive comparison with existing datasets is presented in Tab. 1. In contrast to current manga and story visualization datasets, the proposed MangaZero dataset stands out as being **larger in size**, **newer in source**, **richer in annotations**, **diverse in manga series**, and **varied in panel resolutions**. Compared to the well-known black-and-white manga dataset Manga109 [2], the MangaZero dataset encompasses more manga series published after the year 2000, which inspired its naming. Additionally, MangaZero includes famous se-

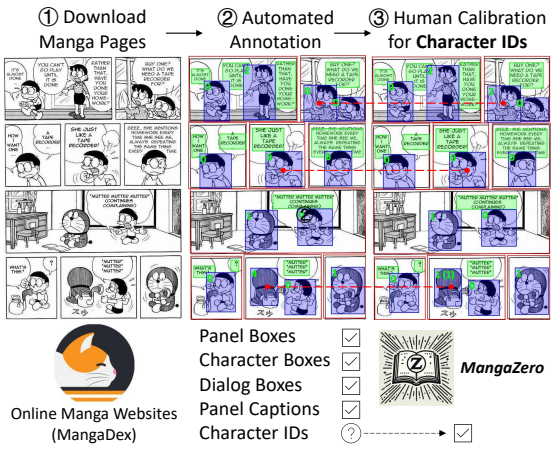


Figure 2. We construct **MangaZero** through three steps: 1) Download manga pages from the internet. 2) Annotate manga panels autonomously with pre-trained models. 3) Human calibration for the character ID annotation.

ries from before the year 2000 that are not featured in Manga109, such as Doraemon (1974).

Construction pipeline. To build our dataset, we first download manga pages from the internet, explicitly sourcing images from MangaDex [1]. It is important to note that all data will be used solely for academic research, not commercial purposes. We select 48 manga series and download up to 1,000 pages per series, resulting in 43,264 double-page images. These images are then annotated using pre-trained models. For manga-specific annotations, including panel bounding boxes, character bounding boxes, character IDs, and dialog bounding boxes, we employ the recent manga understanding model, Magi [30]. It should be noted that character ID labeling is consistent only within individual pages, which is sufficient for achieving coherent character cross-reference. Once the panel bounding boxes are obtained, we utilize LLaVA-v1.6-34B [22] to generate captions for each panel. However, we observe that character ID labeling has relatively low accuracy, which poses a significant challenge for training purposes. To address this, human annotators refine the machine-generated labels, resulting in accurate and clean annotations. Finally, we split 96 pages (2 for each series) as the evaluation set and the remaining 43,168 pages as the training set.

4. Method

In this section, we introduce the architecture of the proposed framework, **DiffSensei**, which generates vivid manga panels with precise character and dialog layout control while adapting the characters' status flexibly.

Motivation. There are two critical problems in customizing objects and layouts in image generation: 1) Preserving the subjects' intrinsic features while avoiding direct copy-paste from source images, and 2) Ensuring reliable layout control with minimal computational cost during both training and inference. To avoid copy-pasting effects, our model converts character image features into tokens, preventing the direct transfer of fine-grained pixel details. Additionally, we integrate an MLLM as a character image feature adapter. The MLLM adapter receives source character features and panel captions as inputs, generating text-compatible target character features. Compared with previous customization work [39], this approach enables text-compatible character encoding and flexible character adaptation to captions. For layout control, we employ lightweight masked encoding techniques for both character and dialog layouts, significantly reducing computational costs compared with previous works [38, 45] while maintaining high accuracy in both training and inference phases. Experiment results in Sec. 5 demonstrate the effectiveness of our design.

Multi-character feature extraction. As illustrated in Fig. 3, we initially extract local image features using CLIP and image-level features from a manga image encoder. These two sets of features are then processed by a feature extractor, which is implemented as a resampler module [47]. This process can be formalized as follows:

$$\mathbf{c}_i = \text{Resampler}([\text{CLIP}(\mathbf{I}), \psi(\mathbf{I})], \mathbf{q}, \mathbf{q}_{void}), \quad (1)$$

where ψ represents the manga image encoder. $\mathbf{q}$ and $\mathbf{q}_{void}$ are trainable query vectors for character and non-character features, respectively. $\mathbf{q}$ re-samples the image features into the U-Net's cross-attention dimension, while $\mathbf{q}_{void}$ guides the cross-attention in regions without characters in the layout. $\mathbf{c}_i \in \mathbb{R}^{B\times((N_c+1)\times N_q)\times C}$ is the output feature for all characters, where B is the batch size, N_c is the maximum number of characters per panel (padded with all-zero features as needed), N_q is the number of query tokens per character, and C is the cross-attention dimension of the U-Net. Through compressing the character images into a few tokens, DiffSensei avoids encoding fine-grained spatial features from reference images into the model [21, 51]. This enables focusing on the characters' semantic representations rather than rigid pixel distributions.

Masked cross-attention injection. We replicate the key and value matrices in each cross-attention layer, creating separate character cross-attention layers. This allows the image query features to attend to text and character cross-attentions independently and combine the results from both attentions. In the character cross-attention, we apply a masked cross-attention injection mechanism to control the layout of each character. Here, each character feature only attends to query features within its designated bounding box region. In areas without characters, query features attend to

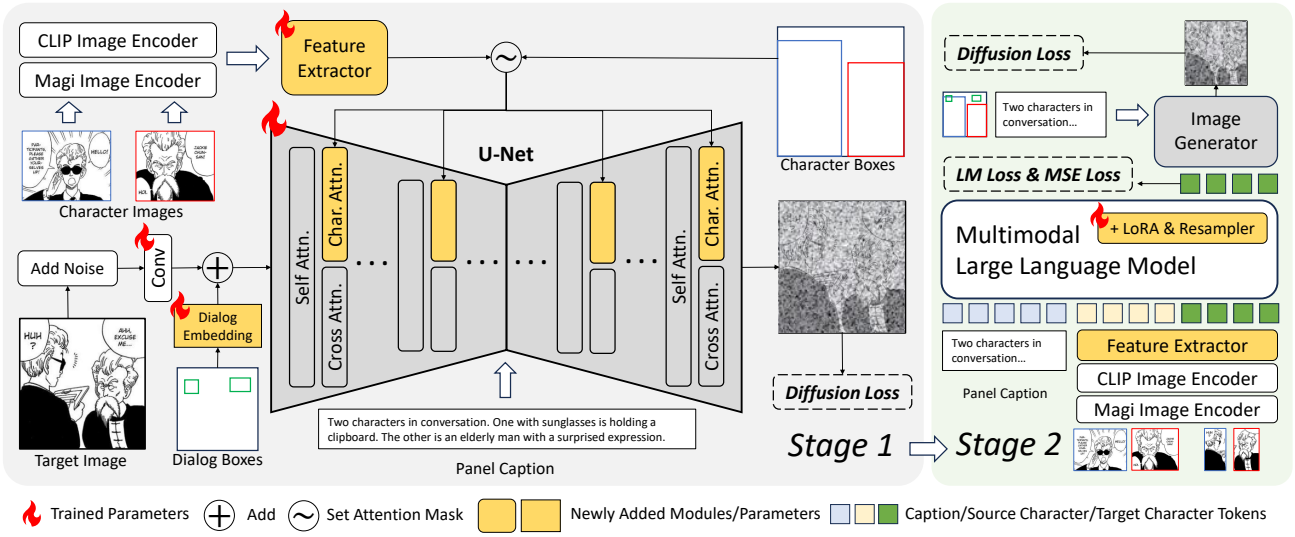


Figure 3. **The architecture of DiffSensei**. In the first stage, we train a multi-character customized manga image generation model with layout control. The dialog embedding is added to the noised latent after the first convolution layer. All the parameters in the U-Net and feature extractor are trained. In the second stage, we finetune LoRA and resampler weights of an MLLM to adapt the source character features corresponding to the text prompt. We use the model in the first stage as the image generator and freeze its weights.

a placeholder vector, $\mathbf{q}_{void}$. This can be formulated as:

$$\begin{aligned}\hat{\mathbf{z}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_t^\top}{\sqrt{d}}\right)\mathbf{V}_t + \\ \alpha \cdot \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_i^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}_i,\end{aligned} \tag{2}$$

where $\mathbf{Q} = \mathbf{z}\mathbf{W}_q$, $\mathbf{K}^t = \mathbf{c}_t\mathbf{W}_k^t$, $\mathbf{V}^t = \mathbf{c}_t\mathbf{W}_v^t$, $\mathbf{K}^i = \mathbf{c}_i\mathbf{W}_k^i$, $\mathbf{V}^i = \mathbf{c}_i\mathbf{W}_v^i$. $\mathbf{Q}$ is the query, $\mathbf{W_q}, \mathbf{W}_k^t, \mathbf{W}_v^t$ are query, key, and value projection matrices for the text cross-attention. $\mathbf{W}_k^t$, $\mathbf{W}_v^t$ are key and value projection matrices for the character cross-attention, initialized from $\mathbf{W}_k^t$ and $\mathbf{W}_v^t$. d is the key dimension. $\mathbf{c}_t$, $\mathbf{c}_i$ are text and character features respectively. $\mathbf{z}$, $\hat{\mathbf{z}}$ are the input and output image features. α is a hyperparameter that controls character attention weight. $\mathbf{M}$ is an attention mask to manage the characters' layout. Its values are defined as follows:

$$\mathbf{M}[i,j] = \begin{cases} 0, & \text{if } j = N_c \text{ and } i \notin \mathbf{B}^c \text{ or} \\ & \quad i \in \mathbf{B}^c[j] \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

where i denotes the position of query tokens, $j \in \{0, 1, ..., N_c\}$ is the character indices. The N_c-th character feature represents the placeholder vector $\mathbf{q}_{void}$. $\mathbf{B}^c[j]$ is the bounding box of the j-th character. The masked attention injection mechanism ensures that each character attends only to its specified bounding box region, while areas without characters attend to the placeholder vector. This technique achieves efficient and precise layout control for each character with minimal computational overhead.

Dialog layout encoding. Panels with dialog are a distinctive feature of manga images. However, most current text-to-image models struggle to generate coherent, readable text [26, 28]. While some recent models can produce stable text, they remain limited in terms of text length [17]. Generating extended text, such as dialogues, continues to pose challenges. Therefore, we propose controlling the layout of dialogs rather than the content itself. In this approach, human artists can manually edit the text within dialog bubbles, leaving image generation to the models. Concretely, we introduce a trainable embedding to represent the dialog layout. The dialog embedding is first expanded to match the spatial shape of the noised latent and then masked with the dialog layout. By summing the masked dialog embedding with the noised latent, we can encode dialog positions within the image generator. This process is expressed as:

$$\hat{\mathbf{z}}_t = \text{Conv}(\mathbf{z}_t) + \text{Expand}(\mathbf{e}_d, \mathbf{z}_t) \cdot \mathbf{M}_d, \tag{4}$$

where $\mathbf{e}_d$ is the trainable dialog embedding, $\mathbf{z}_t$ is the noised latent in time step t, Expand is a function that expands $\mathbf{e}_d$ to the latent shape, and $\mathbf{M}_d$ is the dialog region mask derived from input dialog bounding boxes $\mathbf{B}^d$. The output, $\hat{\mathbf{z}}_t$, serves as a dialog-layout-aware latent representation. This is then input into the U-Net for noise prediction. The dialog embedding effectively encodes the dialog layout, adding minimal computational overhead in space and time.

MLLM as text-compatible character feature adapter. After training the image generator, our model can effectively create manga panels that adhere to specified character appearances and layout conditions. However, the model

often rigidly follows the input character images, lacking flexibility in adjusting expressions, poses, or motions based on panel captions. We propose incorporating MLLM as a text-compatible character feature adapter. This approach allows dynamic modifications to character states based on text prompts. A training sample for MLLM is organized as [panel caption, source character image features, target character image features]. The image features are encapsulated by two special tokens, <IMG>and </IMG>. To achieve this, we compute Language Modeling (LM) Loss on the special tokens to constrain the output format and Mean Squared Error (MSE) Loss to guide the target character features based on the panel caption. To ensure that the edited character features align with the image generator, we further pass the generated features into U-Net's character cross-attention and compute diffusion loss. In this stage, only the LoRA and resampler weights in MLLM are updated. This process can be formalized as follows:

$$\begin{aligned} \hat{\mathbf{h}}, \hat{\mathbf{c}}_i &= \text{MLLM}(T, \phi(\mathbf{c}_i)), \\ \mathcal{L}_{lm} &= \text{LMLoss}(\hat{\mathbf{h}}, T), \\ \mathcal{L}_{mse} &= \text{MSELoss}(\phi'(\hat{\mathbf{c}}_i), \tilde{\mathbf{c}}_i), \\ \mathcal{L}_{diff} &= \mathbb{E}_{\epsilon,t}||\epsilon - \epsilon_\theta(\mathbf{z}_t, T, \hat{\mathbf{c}}_i, \mathbf{B}^c, \mathbf{B}^d, t)||^2, \end{aligned} \tag{5}$$

where T is the text prompt. ϕ and ϕ' denote the input and output resamplers of MLLM, which consist of stacked attention layers to convert the embeddings between inner and outer dimensions. $\hat{\mathbf{h}}$ refers to the MLLM predicted special token embeddings. We calculate LM Loss on it. $\hat{\mathbf{c}}_i$ is the predicted character features. We compute MSE Loss between $\hat{\mathbf{c}}_i$ and $\tilde{\mathbf{c}}_i$, the ground truth target character embedding extracted from the feature extractor. By leveraging the character ID annotations of MangaZero, we can obtain features from the same character across different panels, thus facilitating the training of the MLLM feature adapter. The adapted character feature $\hat{\mathbf{c}}_i$ is then passed to the image generator ϵ, previously trained, to compute a diffusion loss. The total loss for training MLLM is expressed as follows:

$$\mathcal{L} = \lambda_{lm}\mathcal{L}_{lm} + \lambda_{mse}\mathcal{L}_{mse} + \lambda_{diff}\mathcal{L}_{diff}, \tag{6}$$

where λ_{lm}, λ_{mse}, and λ_{diff} are loss weights.

5. Experiments

In this section, we thoroughly evaluate DiffSensei and compare it with baseline models.

5.1. Experimental Settings

Implementation details. The image generator is constructed on top of SDXL [26]. The feature extractor's weights are initialized using the pre-trained IP-Adapter-Plus-SDXL [47], while the MLLM (Multi-modal Large Language Model) is initialized from SEED-X [8]. Other newly introduced parameters, including the LoRA and resampler weights of the MLLM, are initialized randomly. In stage 1, the image generator is optimized using a learning rate of 1e-5. Stage 2 training employs a learning rate of 1e-4 and a LoRA rank of 64 [14]. The optimizer is AdamW [23]. The loss function coefficients λ_{lm}, λ_{mse}, and λ_{diff} are set to 1.0, 6.0, and 1.0, respectively. We train 250k steps for the first stage and 20k for the second stage. The source character images are chosen randomly, with a 50% probability of being from the same page; otherwise, they are selected from the target image. To handle varying image resolutions during training, we adopt the bucket-based approach from prior works [26], grouping images into resolution-specific buckets. For each training batch, images are loaded from the same resolution bucket. The batch size varies between 8 and 64 in stage 1, and between 8 and 128 in stage 2. This dynamic batch sizing is necessary to prevent out-of-memory (OOM) issues, especially when processing large-resolution images. Please see more details in the Appendix.

Evaluation datasets and metrics. We evaluate our model using the MangaZero and Manga109 [2] evaluation sets. Note that the model is only trained on MangaZero. Characters in Manga109 are **unseen** during training, serving as a benchmark for generalization. To assess the generation quality of individual images, we employ autonomous metrics, which include Fréchet Inception Distance score (FID) [13], CLIP image-text similarity (CLIP) [27], DINO image similarity (DINO-I) [24], DINO character image similarity (DINO-C), and the dialog bounding box F1 score (F1 score). The source character images are randomly sampled on the same page. The dialog bounding boxes in the generated images are predicted using Magi [30]. For evaluating the story visualization quality of image sequences, human preference study proves more effective. We recruit human volunteers to choose their preferred story pages from our model's output and baseline models in the MangaZero evaluation set. The evaluation criteria include five key aspects: text-image alignment, style consistency, character consistency, image quality, and overall preference.

Baselines. We select recent advanced story visualization models as our baselines, including StoryDiffusion [53], AR-LDM [25], StoryGen [21], SEED-Story [46], and MS-Diffusion [39]. StoryDiffusion [53] is a training-free method. We directly use an SDXL text-to-image model finetuned on MangaZero for evaluation. Despite that, we re-train other baselines on our dataset for a fair comparison.

5.2. Comparison to Baselines

Quantitative comparison. We quantitatively compare our DiffSensei model with baseline models using automatic evaluation metrics. The results on the MangaZero evaluation set are presented in Tab. 2a. The results highlight that DiffSensei consistently outperforms the baseline mod-

Table 2. **Quantitative comparisons on automatic metrics**. Methods followed by "*" use reference images as input rather than characters. Methods marked by "†" means re-trained with dialog embedding.

(a) Comparison on MangaZero evaluation set.

Method	FID ↓	CLIP ↑	DINO-I ↑	DINO-C ↑	F1 score ↑
AR-LDM* [25]	0.409	**0.257**	0.548	0.507	0.004
StoryGen* [21]	0.411	0.219	0.536	0.488	0.012
SEED-Story* [46]	0.411	0.169	0.416	0.405	0.006
StoryDiffusion* [53]	0.409	0.244	0.461	0.362	0.002
MS-Diffusion† [39]	0.408	0.229	0.610	0.641	0.720
DiffSensei	**0.407**	0.235	**0.618**	**0.651**	**0.727**

(b) Comparison on Manga109 evaluation set.

Method	FID ↓	CLIP ↑	DINO-I ↑	DINO-C ↑	F1 score ↑
AR-LDM* [25]	0.410	**0.254**	0.527	0.491	0.005
StoryGen* [21]	0.414	0.214	0.540	0.493	0.004
SEED-Story* [46]	0.413	0.167	0.442	0.428	0.005
StoryDiffusion* [53]	0.410	0.238	0.442	0.355	0.001
MS-Diffusion† [39]	**0.410**	0.227	0.584	**0.600**	0.601
DiffSensei	**0.410**	0.237	**0.588**	**0.600**	**0.648**

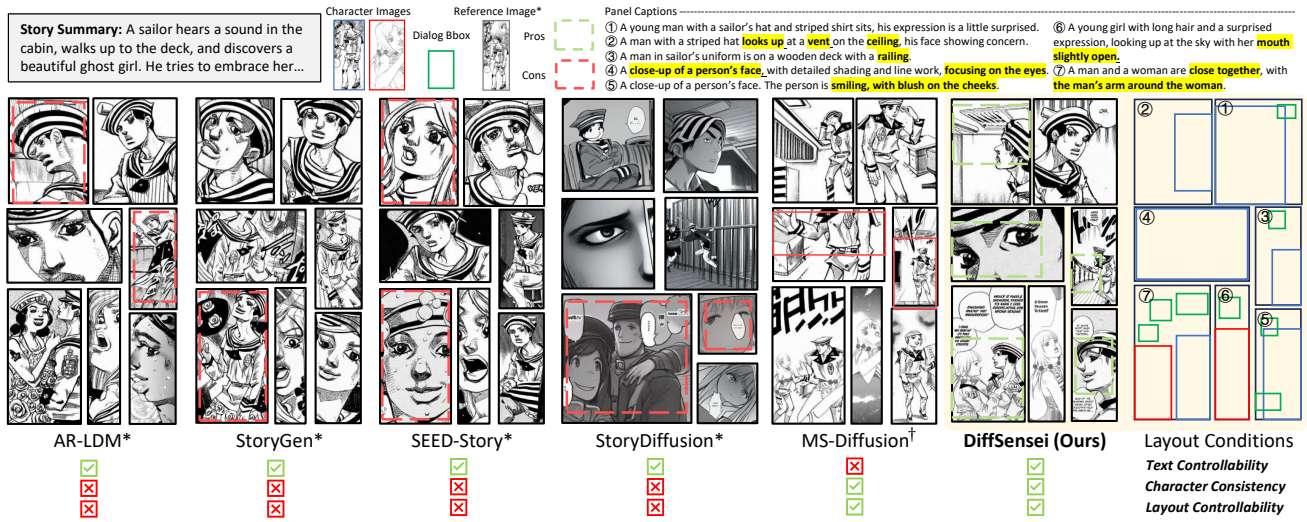


Figure 4. **Qualitative comparison with baselines**. Baselines followed by a "*" use reference images as input rather than character images. Methods marked by "†" means re-trained with dialog embedding. Our model excels at preserving the characters while following the text prompt. Our DiffSensei successively generates highlighted details in panel captions. Better viewed with zoom-in.

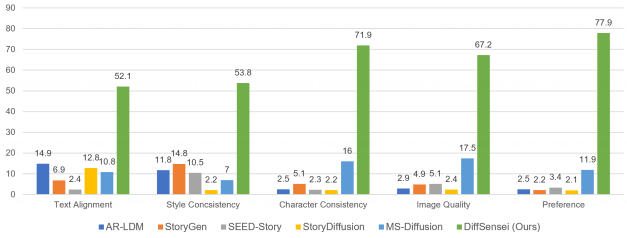


Figure 5. **Human preference study on MangaZero eval set**.

els across five key metrics. Our model improves 0.06 in the CLIP metrics compared to the multi-subject customization baseline, MS-Diffusion [39], which struggles to modify characters' states effectively in response to textual prompts. Furthermore, DiffSensei demonstrates superior image quality and character preservation, as evidenced by higher DINO-I and DINO-C scores. Although AR-LDM [25] achieves higher CLIP metrics, it suffers from poor image quality and lacks the architectural capability to manage multiple characters, resulting in low DINO-C scores. In contrast, our method strikes a balance between maintaining character appearances and adapting to text prompts. We also compare the Manga109 evaluation set with the results shown in Tab. 2b. When using previously unseen characters as inputs, our model continues to outperform the baselines. These results underscore the strong generalization ability of DiffSensei, demonstrating effective adaptation to new characters.

In Fig. 5, we present the results of a human preference study comparing our model to the baselines across several dimensions. Our model receives the highest ratings from human evaluators, particularly in terms of overall preference, character consistency, and image quality. These findings confirm that DiffSensei excels in rendering vivid and engaging manga stories.

Qualitative comparison. Fig. 4 shows a qualitative comparison between DiffSensei and baseline models. The results illustrate that our model significantly outperforms the baselines in generating an entire page of a manga story. SEED-Story [46] employs an MLLM to create captions for each panel, leading to unnatural narrative text and chaotic story generation that fails to form a coherent story. StoryD-

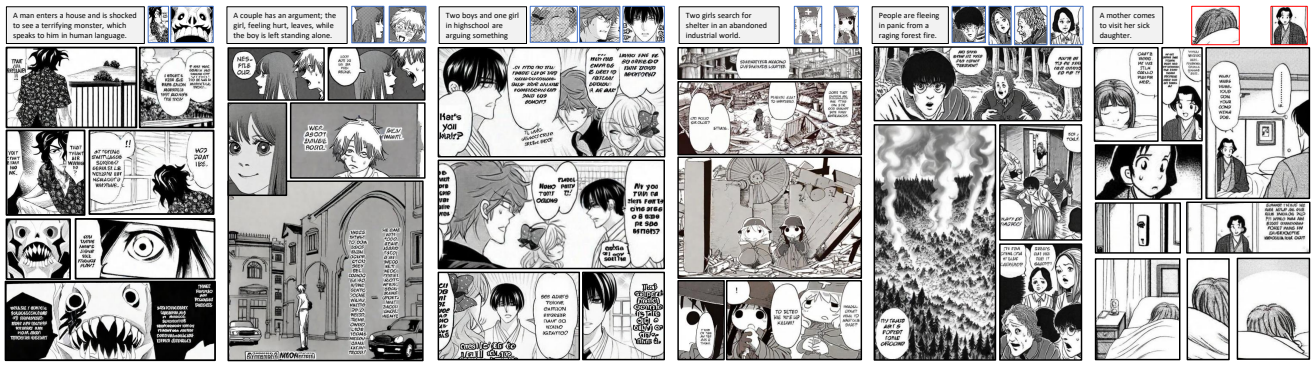


Figure 6. **Qualitative results**. Character images in red boxes are from Manga109 (The rightmost example). Our DiffSensei can generate vivid manga pages in various scenarios. Better viewed with zoom-in. More results can be found in the appendix.

Table 3. **Ablation study**. CM is character masked attention injection. DM is dialog masked encoding. Magi means using Magi [30] image encoder. MLLM means using MLLM for stage 2 training.

CM	DM	Magi	MLLM	FID ↓	CLIP ↑	DINO-I ↑	DINO-C ↑	F1 score ↑
				0.410	0.230	0.593	0.610	0.361
✓				0.411	0.225	0.591	0.637	0.364
✓	✓			**0.407**	0.228	0.600	0.635	0.653
✓	✓	✓		0.408	0.231	**0.618**	0.648	0.718
✓	✓	✓	✓	**0.407**	**0.235**	**0.618**	**0.651**	**0.727**

iffusion [53] is limited to producing fixed-resolution images due to its self-attention sharing mechanism, restricting its ability to generate diverse images. It shows inferior results, probably because the input reference panel has an unbalanced aspect ratio. MS-Diffusion [39] trains with source character images from the target panel and lacks the flexibility to modify characters' states effectively. In contrast, our method excels in text-following, character preservation, and overall story presentation.

5.3. Qualitative Results

Fig. 6 shows several manga pages generated by DiffSensei. The results demonstrate that our method can generate vivid manga panels and provide visually plausible results for the customized manga generation task. Our model can also generalize to unseen characters, as illustrated in the rightmost example, with character images from Manga109 [2] as input. Please see more results in the Appendix.

5.4. Ablation Study

Tab. 3 presents the quantitative ablation study of DiffSensei, where we systematically remove components to assess their impact. Specifically, when the MLLM component, serving as the flexible character feature adapter, is excluded, the CLIP metrics decrease by 1.73%, and the DINO-C score also drops, underscoring its role in enhancing transferring character to the text-derived states. The absence of the Magi [30] image encoder results in a general decline in metrics, particularly in image quality and character similarity, highlighting the importance of the Magi encoder for effectively encoding manga characters. Magi is trained specifically on manga datasets, performing better at preserving manga characters. To investigate alternative methods for encoding character and dialog layout conditions, we experimented with replacing the dialog embedding technique by inputting the Fourier embeddings of dialog bounding boxes into the timestep embedding of SDXL [26]. This modification led to a significant decrease in layout control, evidenced by the F1 score plummeting from 0.653 to 0.364, demonstrating that directly incorporating dialog embeddings into the latent is a superior approach for encoding dialog layouts. Furthermore, we explored adding the Fourier embeddings of character bounding boxes to the character features as an alternative to masked attention injection. This change caused a marked drop in the DINO-C metric, reaffirming the effectiveness of our original masked attention strategy. For comprehensive effect showcases, please see the qualitative ablation study in the Appendix.

6. Conclusion

This paper introduces DiffSensei, a novel framework for multi-character customized story visualization that integrates a diffusion-based image generator with an MLLM as a text-compatible identity adapter. Key innovations include masked attention control for character layout management, dialog layout embedding, and an MLLM-based feature adapter for flexible character customization. Supported by the proposed MangaZero dataset, comprising 43,264 manga pages and 427,147 panels, DiffSensei achieves superior, character-consistent panels that dynamically respond to textual prompts, surpassing existing methods and advancing the field of story visualization.

Acknowledgement. This work is supported by the National Key Research and Development Program of China (No. 2023YFC3807600).

References

[1] Mangadex, 2024. 4

[2] Kiyoharu Aizawa, Azuma Fujimoto, Atsushi Otsubo, Toru Ogawa, Yusuke Matsui, Koki Tsubota, and Hikaru Ikuta. Building a manga dataset "manga109" with annotations for multimedia applications. *TMM*, 2020. 3, 6, 8

[3] Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. *arXiv preprint arXiv:2410.08261*, 2024. 3

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024. 2

[5] Junhao Cheng, Xi Lu, Hanhui Li, Khun Loun Zai, Baiqiao Yin, Yuhao Cheng, Yiqiang Yan, and Xiaodan Liang. Autostudio: Crafting consistent subjects in multi-turn interactive image generation. *arXiv preprint arXiv:2406.01388*, 2024. 1, 2

[6] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. In *ICLR*, 2023. 2, 3

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[8] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 2, 3, 6

[9] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *NeurIPS*, 2024. 2

[10] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, 2018. 2, 3

[11] Yue Han, Jiangning Zhang, Junwei Zhu, Xiangtai Li, Yanhao Ge, Wei Li, Chengjie Wang, Yong Liu, Xiaoming Liu, and Ying Tai. A generalist facex via learning unified facial representation. *arXiv preprint arXiv:2401.00551*, 2023. 2

[12] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. *ECCV*, 2024. 2

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[14] 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. 6

[15] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *NAACL*, 2016. 2

[16] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*, 2024. 2, 3

[17] Black Forest Labs. Announcing black forest labs, 2024. 5

[18] 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 *CVPR*, 2019. 1, 2, 3

[19] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 2

[20] Jian Lin, Xueting Liu, Chengze Li, Minshan Xie, and Tien-Tsin Wong. Sketch2manga: Shaded manga screening from sketch with diffusion models. In *ICIP*, 2024. 2, 3

[21] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *CVPR*, 2024. 1, 2, 3, 4, 6, 7

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 4

[23] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[24] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[25] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *WACV*, 2024. 1, 2, 3, 6, 7

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5, 6, 8

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICCV*, 2021. 6

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5

[29] 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 *CVPR*, 2023. 2, 3

[30] Ragav Sachdeva and Andrew Zisserman. The manga whisperer: Automatically generating transcriptions for comics. In *CVPR*, 2024. 3, 4, 6, 8

[31] Ragav Sachdeva, Gyungin Shin, and Andrew Zisserman. Tails tell tales: Chapter-wide manga transcriptions with character names. *arXiv preprint arXiv:2408.00298*, 2024. 3

[32] Qingyu Shi, Lu Qi, Jianzong Wu, Jinbin Bai, Jingbo Wang, Yunhai Tong, Xiangtai Li, and Ming-Husang Yang. Relationbooth: Towards relation-aware customized object generation. *arXiv preprint arXiv:2410.23280*, 2024. 2

[33] Hao Su, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Jiahe Cui, and Ji Wan. Mangagan: Unpaired photo-to-manga translation based on the methodology of manga drawing. In *AAAI*, 2021. 2, 3

[34] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, 2024. 2, 3

[35] Emanuele Vivoli, Andrey Barsky, Mohamed Ali Souibgui, Artemis LLabres, Marco Bertini, and Dimosthenis Karatzas. One missing piece in vision and language: A survey on comics understanding. *arXiv preprint arXiv:2409.09502*, 2024. 3

[36] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2

[37] Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human effort. *arXiv preprint arXiv:2311.11243*, 2023. 1, 2

[38] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *CVPR*, 2024. 4

[39] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2, 3, 4, 6, 7, 8

[40] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 3

[41] Zhenyu Wang, Aoxue Li, Zhenguo Li, and Xihui Liu. Genartist: Multimodal llm as an agent for unified image generation and editing. In *NeurIPS*, 2024.

[42] Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, et al. Towards language-driven video inpainting via multimodal large language models. In *CVPR*, 2024. 3

[43] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. Towards open vocabulary learning: A survey. *T-PAMI*, 2024. 2

[44] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. In *NeurIPS*, 2024. 2

[45] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *ICCV*, 2023. 4

[46] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*, 2024. 1, 2, 3, 6, 7

[47] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 3, 4, 6

[48] Zhang Yunqian. Ai-driven background generation for manga illustrations: A deep generative model approach. *ORES*, 2024. 2, 3

[49] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *NeurIPS*, 2024. 2, 3

[50] Lvmin Zhang, Xinrui Wang, Qingnan Fan, Yi Ji, and Chunping Liu. Generating manga from illustrations via mimicking manga creation workflow. In *CVPR*, 2021. 2, 3

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 4

[52] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, and Tong Sun. Customization assistant for text-to-image generation. In *CVPR*, 2024. 3

[53] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. In *NeurIPS*, 2024. 1, 2, 6, 7, 8