# Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models

**Fei Shen**[1,2*], **Hu Ye**[2], **Sibo Liu**[2], **Jun Zhang**[2†], **Cong Wang**[2], **Xiao Han**[2], **Wei Yang**[2]

[1]Nanjing University of Science and Technology
[2]Tencent AI Lab

## Abstract

Recent research showcases the considerable potential of conditional diffusion models for generating consistent stories. However, current methods, which primarily generate stories in a caption-dependent manner, often overlook the importance of contextual consistency and the relevance of frames during sequential generation. To address this, we propose a novel **R**ich-contextual **C**onditional **D**iffusion **M**odel**s** (RCDMs), a two-stage approach designed to enhance story generation's semantic consistency and temporal consistency. Specifically, in the first stage, the frame-prior transformer diffusion model is presented to predict the frame semantic embedding of the unknown clip by aligning the semantic correlations between the captions and frames of the known clip. The second stage establishes a robust model with rich contextual conditions, including reference images of the known clip, the predicted frame semantic embedding of the unknown clip, and text embeddings of all captions. By jointly injecting these rich contextual conditions at the image and feature levels, RCDMs can generate semantic and temporal consistency stories. Moreover, RCDMs can generate consistent stories with a single forward inference compared to autoregressive models. Our qualitative and quantitative results demonstrate that our proposed RCDMs outperform in challenging scenarios.

## Introduction

Story visualization (Li et al. 2019; Rahman et al. 2023; Pan et al. 2024) aims to depict a continuous narrative through multiple captions and reference clips. It has profound applications in game development and comic drawing. Due to the technological leaps in generative models such as generative adversarial network (GAN) (Creswell et al. 2018; Qiao et al. 2019; Zhang et al. 2019; Ding et al. 2020; Hong et al. 2022; Li et al. 2023a,b, 2024d) and diffusion model (Ramesh et al. 2022; Long et al. 2024; Zhang, Rao, and Agrawala 2023; Saharia et al. 2022a; Shen et al. 2023; Ye et al. 2023), text-to-image synthesis methods (Zhang et al. 2021; Xie et al. 2024b,c,a; Ramesh et al. 2021; Zhang et al. 2023; Yang et al. 2023; Li et al. 2024c) can now generate visually faithful images through text descriptions. However, given multiple captions to generate a continuous story with style and temporal consistency still poses significant challenges.

Existing methods typically employ autoregressive generation and can be broadly classified into GAN-based (Li et al. 2019; Maharana, Hannan, and Bansal 2021; Li 2022) and diffusion model-based (Pan et al. 2024; Maharana, Hannan, and Bansal 2022; Rahman et al. 2023; Shen and Elhoseiny 2023). GAN-based methods typically comprise a text encoder, image generator, image separation, and story discriminator. These components work together to maintain the consistency of the entire sequence of images. However, the images generated by these methods often display distorted objects, mismatched semantics, and localized blurring, especially when creating images from complex scene descriptions. Subsequently, GAN-based methods (Maharana, Hannan, and Bansal 2021; Maharana and Bansal 2021) progressively focus on generating more consistent images by improving the performance of the text encoder, such as caption enhancement (Maharana, Hannan, and Bansal 2021), structured text parsing (Maharana and Bansal 2021), and ID attention mechanism (Chen et al. 2022). While these methods (Maharana, Hannan, and Bansal 2021; Maharana and Bansal 2021) can generate images that satisfy the requirements of character consistency, they often struggle to maintain a consistent style and capture realistic scene details. Furthermore, since the adversarial nature of the min-max objective, GAN-based methods can be prone to unstable training dynamics, which limits the diversity of the stories.

Advanced diffusion models like Imagen (Saharia et al. 2022b) and Stable Diffusion (Rombach et al. 2022a) have recently demonstrated unprecedented text-to-image synthesis capabilities. Story visualization based on diffusion models (Pan et al. 2024; Rahman et al. 2023; Shen and Elhoseiny 2023) generates better consistent images through a multi-step denoising process, using the features of the current caption and historical frames as conditions. However, these methods (Pan et al. 2024; Maharana, Hannan, and Bansal 2022; Rahman et al. 2023; Shen and Elhoseiny 2023) only consider clip information at the feature level due to the injection of the current caption and the known clip into the model via the inherent cross-attention module, overlooking the frame information of known at the image level and text information of the unknown clip. Besides, from Figure 1 (a), GAN-based and diffusion model-based autoregressive gen-

[*]Work done during an internship at Tencent Al Lab.
[†]Corresponding Author.

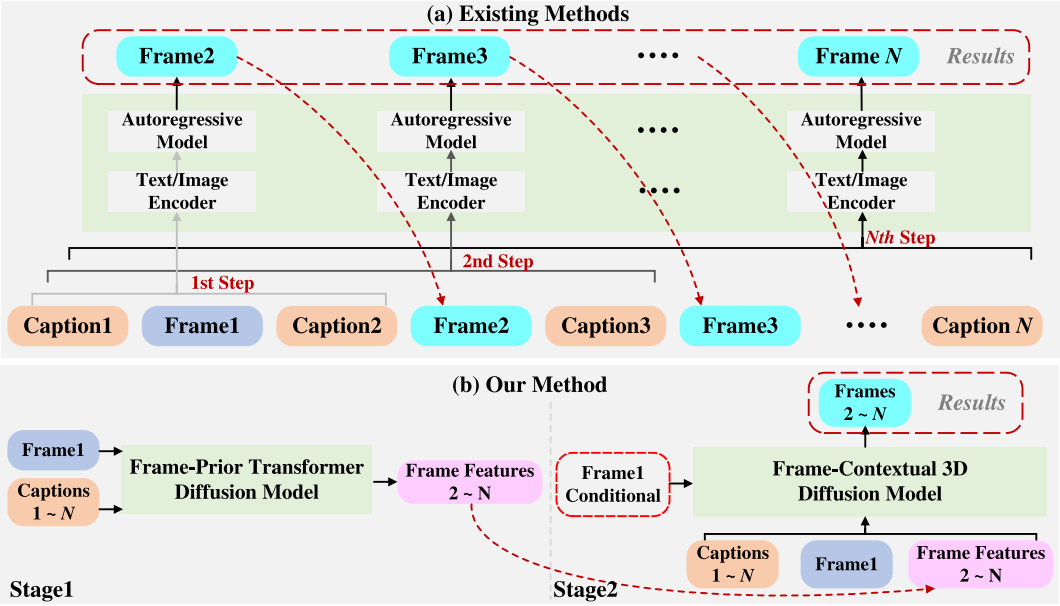


Figure 1: (a) Existing autoregressive methods rely on the current caption, resulting in weak conditioning and reduced story consistency. (b) RCDMs predict frame-contextual features and infuse both image- and feature-level context for coherent story generation in a single forward pass.

eration methods (Pan et al. 2024; Xiao et al. 2023; Lee et al. 2022) depend on the current caption for frame-by-frame forward generation. This reliance can lead these methods to easily overlook the rich contextual information within the captions and the consistency of the story.

We propose **R**ich-contextual **C**onditional **D**iffusion **M**odel**s** (RCDMs) to tackle the issues above through two stages, as shown in Figure 1 (b). Firstly, we propose a frame-prior transformer diffusion model to predict the semantic embeddings of frames within an unknown clip. This prediction task is significantly less complex than directly generating consistent a story. It enables the model to focus solely on the task at the semantic feature level, thereby circumventing the rigorous consistency requirements associated with story generation at the image level. The frame-prior transformer diffusion model takes the frames of the known clip and all captions as conditions, leveraging a combination of multiple cascaded transformer blocks and frame attention blocks to predict the frame semantic embedding of the unknown clip. In the second stage, instead of over-reliance on caption conditions, we integrate the text embeddings of all captions, the frame semantic features of the unknown clip, and the images of the known clip to serve as a rich contextual guide for generating frames in the unknown clip. Based on the above rich-contextual conditional, we devise a frame-contextual 3D diffusion model to generate consistent stories by jointly infusing conditions at both the image and feature levels. Besides, RCDMs generates consistent stories with a single forward inference compared to autoregressive models. The main contributions are summarized as follows:

- We propose a frame-prior transformer diffusion model that, by using the frames from the known clip and all captions as conditions, predicts the semantic embeddings of frames in an unknown clip, thereby providing a rich semantic feature for the next stage.
- We devise a frame-contextual 3D diffusion model that jointly infuses image-level and feature-level rich-contextual conditional to generate stories with stylistic and temporal consistency.
- We conduct comprehensive experiments on two datasets to demonstrate the competitive performance of proposed RCDMs. Additionally, we perform a user study to evaluate the superiority of RCDMs qualitatively.

## Related Work

**Text-to-Image Synthesis.** Recent advancements in text-to-image synthesis have primarily centered around generative adversarial networks (GANs) and diffusion models. MirrorGAN (Qiao et al. 2019) enhances image diversity and semantic consistency through redescription, using both local words and global captions as conditions. AttnGAN (Xu et al. 2018) refines details by matching image sub-regions with relevant words from captions. StackGAN (Zhang et al. 2017) adopts a two-stage process: first sketching objects based on captions, then refining them. XMC-GAN (Zhang et al. 2021) captures caption-image correspondence through attention self-modulation and contrastive discrimination. However, GAN-based methods (Duan et al. 2024; Chen et al. 2023, 2024; Li et al. 2024b,a; Zhang et al. 2021; Yuan et al. 2024b,a) often struggle with hyperparameter tuning and loss of scene details. In contrast, diffusion models (Song et al. 2020; Wang et al. 2024b; Ho, Jain, and Abbeel 2020; Rombach et al. 2022b; Wang et al. 2024a;

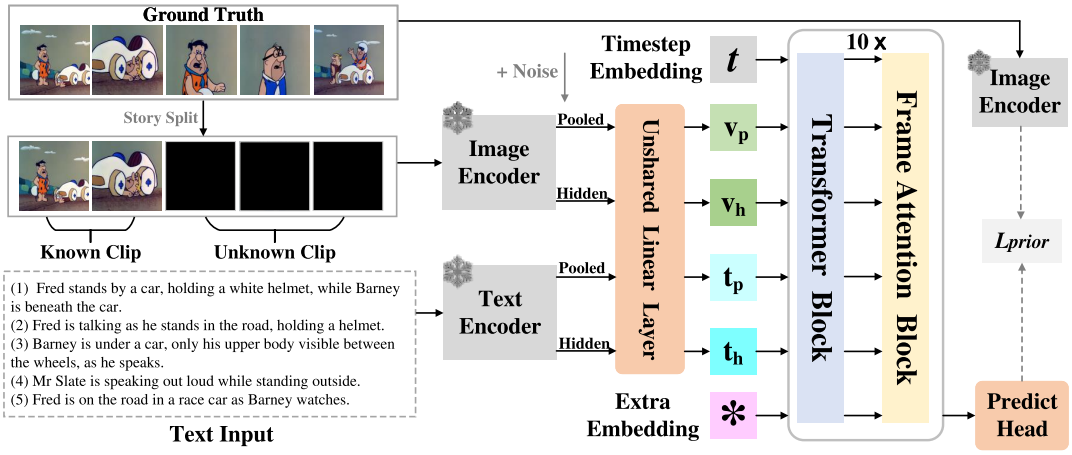


Figure 2: Illustration of the frame-prior transformer diffusion model, which predicts frame semantic embeddings of unknown clips by aligning semantic correlations between captions and frames of known clips.

Shen et al. 2024; Shen and Tang 2024) avoid mode collapse and unstable training, generating more diverse images. DALL-E2 (Ramesh et al. 2022) combines autoregressive CLIP embedding generation with diffusion-based decoding for text-consistent images. Imagen (Saharia et al. 2022a) leverages a large transformer language model for enhanced image-text alignment and fidelity. Stable Diffusion (Rombach et al. 2022b) applies diffusion models in latent space, balancing simplicity and detail preservation. Despite these advancements, current text-to-image methods (Saharia et al. 2022a; Luo et al. 2024; Wang, Li, and Cui 2024; Wang, Cui, and Li 2023; Zhang et al. 2024; Zhou et al. 2024; Ge et al. 2023; Wu et al. 2023) often focus on individual image-text alignment, neglecting style and temporal consistency crucial for story visualization.

**Story Visualization.** StoryGAN (Li et al. 2019), a pioneer in story visualization, introduced a sequential GAN framework with a context encoder for tracking story flow and a story-level discriminator. DuCo-StoryGAN (Maharana, Hannan, and Bansal 2021) enhances story-image semantic consistency through dual learning. VLC-StoryGAN (Maharana and Bansal 2021) uses a Transformer-based recursive architecture with structured text input and common sense information to align with story structure. Word-Level SV (Li 2022) improves image quality and story consistency through feature fusion and spatial attention. VP-CSV (Chen et al. 2022) ensures character consistency via a two-stage framework predicting character and remaining tokens. Recent developments in story visualization (Pan et al. 2024; Ahn et al. 2023; Shen and Elhoseiny 2023), especially with diffusion models, have further advanced the field. StoryDALL-E (Maharana, Hannan, and Bansal 2022) employs an autoregressive Transformer for story visualization with fine-tuning and prompt-based adjustments. AR-LDM (Pan et al. 2024) enhances story consistency by aligning historical captions and generated images through an autoregressive latent diffusion model. Similarly, Story-LDM (Rahman et al. 2023) captures character and background context through a visual memory module. However, these autoregressive methods, whether GAN or diffusion-based, often rely on captions for guidance, leading to weaker conditioning and reduced story consistency.

## Method

The proposed **R**ich-contextual **C**onditional **D**iffusion **M**odel**s** (RCDMs) aims to generate consistent stories by jointly infusing rich contextual conditions at both the image and feature levels. In this section, we introduce the following two aspects: frame-prior transformer diffusion model and frame-contextual 3D diffusion model.

### Frame-Prior Transformer Diffusion Model

In the first stage, we propose a frame-prior transformer diffusion model to predict the frame semantic embeddings of unknown clips by aligning the semantic correlations between the captions and frames of known clips. As shown in Figure 2, it is composed of a frozen image encoder, a frozen text encoder, an unshared linear layer, and a stack of multiple transformer blocks and frame attention blocks. Here, we utilize the pooled representation extracted from the CLIP image encoder as the frame semantic embeddings for unknown clips. Our choice is inspired by the capability of CLIP, which is trained on a large-scale dataset of image-text pairs through contrastive learning. This enables it to encapsulate a rich variety of image content and stylistic information, pivotal in steering the subsequent process of story synthesis.

Specifically, we first split the ground truth into known and unknown clips and introduce noise into the unknown clips. Subsequently, all clips (known and unknown) and all captions are fed into the frozen image encoder and the frozen text encoder, respectively. We then use an unshared linear layer to obtain the pooled visual representation $v_p$ and hidden visual representation $v_h$ of the image and pooled textual representation $t_p$ and hidden textual representation $t_h$ of the caption. For consecutive frames, we input them as a 4D tensor $x \in \mathbb{R}^{b \times f \times n \times d}$, where $b$, $f$, $n$, and $d$ represent the

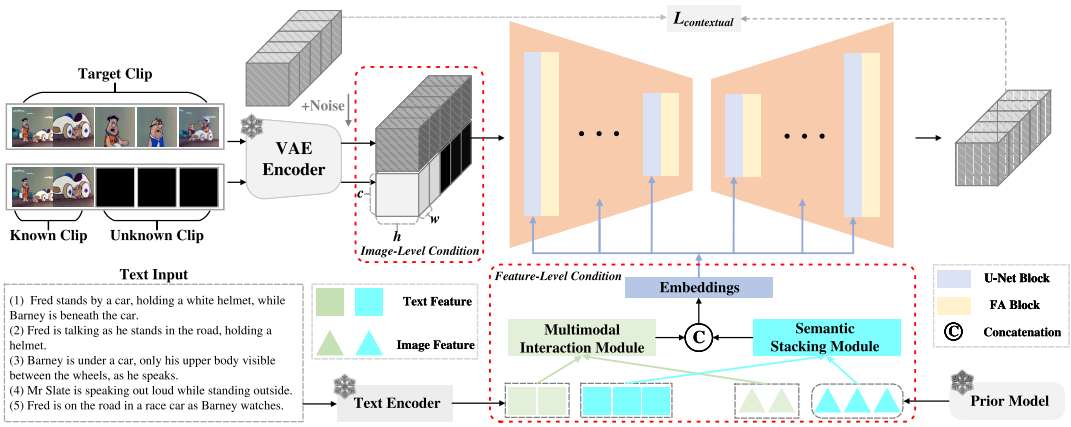


Figure 3: Overview of the frame-contextual 3D diffusion model. The frame-contextual 3D diffusion model infuses both image-level and feature-level context information to generate stories with stylistic and temporal consistency. Prior model denotes frame-prior transformer diffusion model.

batch size, temporal length, token length and each token dimension, respectively. When the internal feature map passes through the transformer block, the temporal length $f$ is reshaped to the batch size $b$ and ignored, allowing the model to process each frame independently. We reshape the feature map back into a 4D tensor after the transformer block, frame attention block reshape $n$ into $b$ to learn and maintain the temporal consistency of the story, and then reshape it back after the module while ignoring the temporal length. Inspired by (Blattmann et al. 2023; Guo et al. 2023), the frame attention block consists of several self-attention modules. This allows the model to guide self-attention along the temporal length $f$, effectively capturing the dynamic content within the narrative. Besides, we add an extra embedding to represent the unnoised frame semantic embedding of the known clip to be predicted.

Following unCLIP (Ramesh et al. 2022), the frame-prior transformer diffusion model is trained to predict the unnoised frame semantic embedding directly rather than the noise added to the frame embedding. The training loss $L_{prior}$ of frame-prior transformer diffusion model $x_\theta$ is defined as follows,

$$L_{\text{prior}} = \mathbb{E}_{x_0,\epsilon,v_p,v_h,t_p,t_h,t} \left\| x_0 - x_\theta \left( x_t, v_p, v_h, t_p, t_h, t \right) \right\|^2 . \tag{1}$$

Once the model learns the conditional distribution, the inference is performed according to Eq. 2, as follows, $w$ is the guidance scale.

$$\begin{aligned} \hat{x}_\theta \left( x_t, v_p, v_h, t_p, t_h, t \right) = & w x_\theta \left( x_t, v_p, v_h, t_p, t_h, t \right) \\ & + (1 - w) x_\theta \left( x_t, t \right) . \end{aligned} \tag{2}$$

## Frame-Contextual 3D Diffusion Model

In the second stage, we propose a frame-contextual 3D diffusion model that utilizes a variety of rich contextual conditions, including reference images from the known clip, the anticipated frame semantic embedding from the unknown clip, and text embeddings from all captions, to generate consistent stories. Therefore, our method can generate style and temporal consistency stories by jointly injecting these rich contextual conditions at the image and feature levels. From Figure 3, the frame-contextual 3D diffusion model comprises a frozen VAE, a model stacked with multiple U-Net blocks and frame attention (FA) blocks, a multimodal interaction module, and a stacked semantic module. Here, the design of the FA block is the same as in the previous stage, and focus on the correlation between frames to maintain consistency across frames.

**Image-Level Condition.** Since the VAE (Kingma, Welling et al. 2019) enables almost lossless reconstruction, introducing known clip conditions at the image level can steer the story continuity, which has been neglected in previous studies. Specifically, similar to the previous stage, we first divide the ground truth into known and unknown clips and directly mask the unknown clip. The ground truth, known clip, and masked unknown clip are all fed into the frozen VAE encoder to extract latent space features. We then concatenate the latent space features of the known clip and masked unknown clip along the width dimension, and simultaneously concatenate them with the latent space features of the ground truth along the channel dimension. Moreover, if the images in the known clip have 0-pixel values similar to the mask, it can easily mislead the model into thinking these areas need to be generated. Therefore, we introduce a single-channel marker symbol that matches the width and height of the concatenated features (omitted in the figure). We use 0 and 1 to represent masked and unmasked pixels, respectively. This approach helps to reduce model confusion and ensures accurate identification of the generation area.

**Feature-Level Condition.** Existing methods depend solely on the current caption's text embeddings, lacking in maintaining the contextual consistency of the overall narrative. Contrarily, we incorporate the frame semantic embedding of the unknown clip obtained from the previous stage and the text embeddings of all captions as extra feature-level conditions. These rich contextual conditions facilitate the generation of consistent narratives. Furthermore, to enhance the

features of the text and frames in both known and unknown segments, we separately introduce a multimodal interaction module and a semantic stacking module. Specifically, we first employ frozen text and image encoders (omitted in the figure) to extract the text embeddings from all captions, and the image embeddings from known clips, respectively. We further divide text embeddings $F_t \in \mathbb{R}^{f\times n\times d}$ along the temporal dimension into $F_t^k \in \mathbb{R}^{f^k\times n\times d}$ and $F_t^u \in \mathbb{R}^{f^u\times n\times d}$ based on known/unknown clips to align the text and image modalities, where $f = f^k + f^u$. For known clip, we first feed the text embeddings $F_t^k$ and image embeddings $F_v^k \in \mathbb{R}^{f^k\times n\times d}$ of the known clip into a multimodal interaction module. The multimodal interaction module comprises two projection layers and a multi-head cross-attention module. The two projection layers are respectively used to project the embeddings of images and text onto the same dimension. Then, the projected text and image embeddings are then fed into the multi-head cross-attention module to obtain the interaction features $F_i^k \in \mathbb{R}^{f^k\times n\times d}$ of the known clip. Considering the discrepancy in the number of tokens and dimensions between the frame semantic embedding and the text embedding, we introduce a semantic stacking module specifically for unknown clips. This module aims to enhance the feature-level semantic information of the unknown clip. The semantic stacking module is composed of a projection layer and a multi-head cross-attention module. The projection layer's role is to convert the text embedding of the unknown clip into the same feature dimension as the image embedding of the unknown clip, which was obtained from the previous stage. Assume that $F_v^u \in \mathbb{R}^{f^u\times 1\times d}$ and $F_t^u \in \mathbb{R}^{f^u\times n\times d}$ respectively are image embedding and projected text embedding of the unknown clip. To align text and image modal, we first obtain the extracted pooled representation $e^g \in \mathbb{R}^{f^u\times 1\times d}$ of unknown clip captions. Then, the pooled representation $F_v^u$ of the unknown clip is fed into the multi-head cross-attention module to obtain the interaction features of the unknown clip. These are then stacked with the hidden representation of the unknown clip along the length dimension to obtain the stacked features $F_s^u \in \mathbb{R}^{f^u\times n\times d}$ of the unknown clip. Finally, the interaction features $F_i^k$ and the stacked features $F_s^u$ are concatenated along the temporal dimension and fed into the U-Net block via inherent cross-attention mechanism in diffusion models.

The loss function $L_{\text{contextual}}$ of frame-contextual 3D diffusion model according to Eq. 3, as follows. Here, $F_I$ denotes the feature of the image-level conditional.

$$L_{\text{contextual}} = \mathbb{E}_{x_0,\epsilon,F_I,F_i^k,F_s^u,t}\|\epsilon - \epsilon_\theta\big(x_t, F_I, F_i^k, F_s^u, t\big)\|^2. \tag{3}$$

In the inference stage, we also use classifier-free guidance according to Eq. 4.

$$\begin{aligned}\hat{x}_\theta\left(x_t, v_p, v_h, t_p, t_h, t\right) = w\, x_\theta\left(x_t, v_p, v_h, t_p, t_h, t\right) \\ + (1-w)\, x_\theta\left(x_t, t\right).\end{aligned} \tag{4}$$

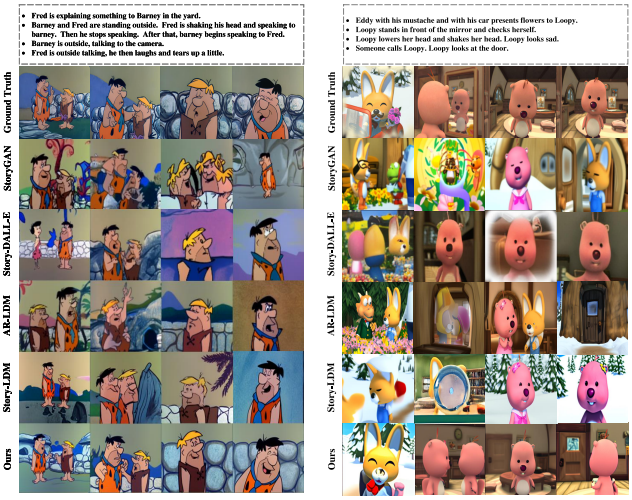


Figure 4: Qualitative comparisons with several state-of-the-art models on the FlintstonesSV and PororoSV datasets. Please see Appendix for more examples.

## Experiments

**Datasets.** We conduct experiments on the FlintstonesSV (Maharana and Bansal 2021) and PororoSV (Li et al. 2019) datasets. The former contains 20,132 training sequences and 2,309 testing sequences, encompassing 7 main characters. PororoSV includes 10,191 training sequences and 2,208 testing sequences, covering 9 main characters. Following (Pan et al. 2024; Rahman et al. 2023; Shen and Elhoseiny 2023), for story visualization, we designated the first frame as the source frame and generated the remaining four based on this source frame.

**Metrics.** We conduct a comprehensive evaluation of the model, considering both objective and subjective metrics. Objective indicators include classification accuracy of characters (Char-Acc) and F1-score of characters (Char-F1), both extracted using InceptionV3. Additionally, we also consider the fréchet inception distance (FID) (Heusel et al. 2017) score. This metric provides a quality assessment by comparing the distribution of feature vectors derived from both real and generated images. In contrast, subjective assessments prioritize user-oriented metrics (Shen and Elhoseiny 2023), including the percentage of visual quality, text-image relevance, and temporal consistency.

**Implementations.** We perform our experiments on 8 NVIDIA V100 GPUs. Our configurations can be summarized as follows, (1) In the frame-prior transformer diffusion model, there are 10 layers of cascaded transformer and frame attention blocks, and the width of each transformer block is 2048. For the frame-contextual 3D diffusion model, we use the pretrained Stable Diffusion V1.5 [1] and modify the first convolution layer to adapt additional conditions. (2) We employ the AdamW optimizer with a fixed learning rate of $1e^{-5}$ in all stages. (3) Following (Pan et al. 2024; Rahman et al. 2023), we train our models using images of sizes

[1] https://huggingface.co/runwayml/stable-diffusion-v1-5

| Datasets | Methods | FID (↓) | Char-Acc (↑) | Char-F1 (↑) |
|---|---|---|---|---|
| FlintstonesSV (Maharana and Bansal 2021) | LDM (Rombach et al. 2022a) | 82.53 | 9.17 | 22.68 |
| | StoryGAN (Li et al. 2019) | 74.63 | 16.57 | 39.68 |
| | Story-DALL-E (Maharana, Hannan, and Bansal 2022) | 26.49 | 55.19 | 73.43 |
| | Story-LDM (Rahman et al. 2023) | 24.24 | 57.19 | 76.59 |
| | AR-LDM (Pan et al. 2024) | 19.28 | 62.58 | 79.25 |
| | **RCDMs (Ours)** | **14.96** | **78.44** | **85.51** |
| PororoSV (Li et al. 2019) | LDM (Rombach et al. 2022a) | 64.52 | 4.31 | 12.74 |
| | StoryGAN (Li et al. 2019) | 49.27 | 9.34 | 18.59 |
| | CP-CSV (Chen et al. 2022) | 40.56 | 10.03 | 21.78 |
| | DuCo-StoryGAN (Maharana, Hannan, and Bansal 2021) | 37.15 | 13.97 | 38.01 |
| | Story-DALL-E (Maharana, Hannan, and Bansal 2022) | 35.90 | 27.14 | 42.45 |
| | Story-LDM (Rahman et al. 2023) | 26.64 | 29.19 | 47.56 |
| | AR-LDM (Pan et al. 2024) | 17.40 | 35.18 | 55.29 |
| | **RCDMs (Ours)** | **16.25** | **41.48** | **59.03** |

Table 1: Quantitative comparison of the proposed RCDMs with several SOTA models.

$512 \times 512$ for FlintstonesSV and PororoSV dataset. (4) We employ a data augmentation strategy of dropping images in all two stages, with the drop count ranging from 0 to 5. We substitute the dropped images with black images. (5) In the inference stage, we use the DDIM (Ho, Jain, and Abbeel 2020) sampler with 20 steps and set the guidance scale $w$ to 2.0 for RCDMs on all stages.

## Quantitative and Qualitative Results

**Quantitative Results.** As shown in Table 1, firstly, since LDM (Rombach et al. 2022a) generates each image based solely on individual captions, it performs significantly worse than all other methods on three metrics. Secondly, compared to GAN and diffusion model methods, our approach outperforms other models on all three metrics in FlintstonesSV. For example, compared to StoryGAN (Li et al. 2019), which employs story dynamic tracking, RCDMs score 61.87% and 45.83% higher on Char-Acc and Char-F1 metrics, respectively. This demonstrates the superiority of proposed RCDMs in understanding story details through all captions and then generating all story images at once, as opposed to StoryGAN, which uses an autoregressive approach to understand captions frame by frame. Lastly, compared to AR-LDM (Pan et al. 2024), which also relies on a diffusion model, RCDMs perform better on the FID metric. Even though AR-LDM already scores well, RCDMs show better performance, indicating that injecting more semantic information through the first-stage model can enrich the generation of image details. Moreover, on Char-Acc and Char-F1 metrics, RCDMs significantly outperform AR-LDM. This is due to introducing more semantic information at the feature level and additional context at the image level.

The comparison results for PororoSV are summarized in Table 1. Notably, consistent with the trend presented in the FlintstonesSV dataset, proposed RCDMs outperform all methods, achieving the best FID, Char-Acc, and Char-F1. Specifically, compared to the best-performing GAN method, i.e., DuCo-StoryGAN, which enhances the current caption to improve text semantic understanding, RCDMs inject context by understanding the complete story semantics. Similarly, compared to AR-LADM, we surpass it by 6.30% and 3.74% on Char-Acc and Char-F1. These results indicate that the simultaneous use of a frame-prior transformer diffusion model to obtain frame semantic information of the unknown clip and the injection of image-level conditions are crucial for understanding and generating stories.

**Qualitative Results.** As some methods have yet to be open-sourced, we qualitatively compared RCDMs with StoryGAN (Li et al. 2019), Story-DALL-E (Maharana, Hannan, and Bansal 2022), AR-LDM (Pan et al. 2024), and Story-LDM (Rahman et al. 2023) on the FlintstonesSV and PororoSV datasets. As shown in Figure 4, several conclusions can be drawn from the results: (1) RCDMs significantly outperform other SOTA methods regarding image quality. For example, on the FlintstonesSV dataset, Story-DALL-E, AR-LDM, and the second frame of Story-LDM, and the third frame of StoryGAN all exhibit scenes and character limbs that do not match. (2) Regardless of whether it is the FlintstonesSV dataset or the PororoSV dataset, our proposed RCDMs perform best regarding character consistency. For example, StoryGAN performs poorly in terms of character consistency on text-image pairs, such as the second and third frames on FlintstonesSV dataset, and the first and fourth frames on PororoSV dataset. A similar situation also occurs on Story-DALL-E. (3) Only our method can generate story images that reasonably align with the text on complex micro-actions and expressions. For example, the text prompt for the fourth frame on the FlintstonesSV dataset is 'laughs and tears up a little,' and the third frame on the PororoSV dataset is 'look sad'. This can be attributed to our method's ability to enhance the semantic consistency of image-text pairs. (4) Regarding visual consistency, although Story-LDM integrates an attention-memory module to handle context, it cannot produce a consistent style scene in a complete story. In contrast, the results of RCDMs have pleasing visual effects and the ability to maintain temporal consistency, primarily due to RCDMs injecting rich contextual conditions, including reference images of the known clip, the predicted frame semantic embedding of the unknown clip, and text embeddings of all captions at both the feature and image levels. In summary, our method can always produce more realistic and consistent story images,

| Settings | Components | | | | FlintstonesSV | | |
|---|---|---|---|---|---|---|---|
| | Stage1 | Stage2 | | | FID (↓) | Char-Acc (↑) | Char-F1 (↑) |
| | | IC | MIM | SSM | | | |
| LDM (Rombach et al. 2022a) | - | - | - | - | 82.53 | 9.17 | 22.68 |
| B0 | ✓ | ✗ | ✗ | ✗ | 21.81 | 56.44 | 70.32 |
| B1 | ✓ | ✓ | ✗ | ✗ | 19.46 | 61.88 | 78.03 |
| B2 | ✓ | ✓ | ✓ | ✗ | 18.36 | 72.53 | 82.06 |
| B3 | ✓ | ✓ | ✗ | ✓ | 17.94 | 73.58 | 82.87 |
| B4 | ✗ | ✓ | ✓ | ✓ | 16.51 | 75.73 | 83.96 |
| **Ours** | ✓ | ✓ | ✓ | ✓ | **14.96** | **78.44** | **85.51** |

Table 2: Ablation study on FlintstonesSV dataset. Here, IC stands for image-level conditional. MIM and SSM, respectively, denote the multimodal interaction module and semantic stacking module.

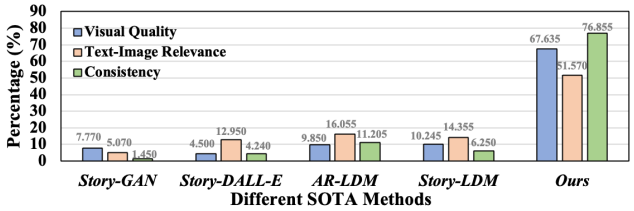


Figure 5: Results of user study. Higher values indicate better performance. RCDMs outperform others across all metrics.

proving that RCDMs bring significant advantages by introducing rich context conditions.

**User Study.** The above quantitative and qualitative comparison results demonstrate the substantial advantages of our proposed RCDMs in generating results. However, the task of synthesizing story visualization is often human perception-oriented. Therefore, we also conduct a user study involving 100 volunteers with computer vision backgrounds. The volunteers are asked to choose which method is better regarding visual quality, text-image relevance, and style/temporal consistency in the generated story.

As shown in Figure 5, the higher the score in the three indicators of this study, the better the performance. RCDMs offer commendable performance on all three fundamental indicators. For example, the text-image relevance and consistency scores of RCDMs are 51.570% and 76.855%, respectively, which are nearly 35.515% and 65.650% higher than the second-best model. This demonstrates the significant advantages of RCDMs in multimodal semantic understanding and consistency due to our conditional injection at both the image and feature levels. In addition, our visual quality score is 67.635%, indicating that participants prefer our method, demonstrating better story visualization quality.

## Ablation Study

We further devise several variants to demonstrate the efficacy of each module proposed in this study. All these variants belong to the RCDMs framework but encompass different configurations. **B0** denotes the sole use of the frame-prior transformer diffusion model, devoid of image-level conditions, with feature-level conditions directly infused into the inherent cross-attention of SD using caption features. **B1** built on the foundation of B0, incorporates image-level conditions. **B2**, an extension of B1, additionally employs a multimodal interaction module. **B3** also building upon B1, further utilizes a semantic stacking module. **B4** indicates that there is no stage1 prior model and in the SSM without stage1, the QKV of attention all originate from the features of the known clips.

From Table 2, firstly, compared to LDM, B0, which adopts the frame attention module, can significantly improve the consistency of characters and backgrounds. Subsequently, when the image-level condition setting is added, B1 is 5.44% and 7.71% higher than B0 on Char-Acc and Char-F1. This demonstrates that the injection of known clip images can enhance the model's contextual information. Secondly, when the MIM and SSM are introduced based on B1, B2 and B3 are 10.65% and 11.70% higher than B1 on Char-Acc. At this point, B2 and B3 have achieved highly competitive performance on the FlintstonesSV dataset. These results indicate that they also contribute constructively to the success of our RCDMs. RCDMs outperform the B4 setting, highlighting the importance of acquiring prior features in the first stage of our method. Finally, when the frame semantic embeddings of the unknown clip predicted by the frame-prior transformer diffusion model are added, FID, Char-Acc, and Char-F1 all improve better, especially FID. This shows that the frame-prior model can better help RCDMs generate stories with semantic and temporal consistency.

## Additional Results

**Branching Storyline.** To further validate RCDMs ability to produce branching narratives, we conducted an additional experiment using one reference image and two distinct sets of captions as story descriptions. To test the model's capacity for maintaining narrative coherence, we included pronouns like "they." From Figure 6, RCDMs generated two diverging story sequences that remained consistent in style and progression. Notably, when encountering the pronoun "they," the model effectively inferred two distinct characters, preserving coherence with the previously parsed storylines.

**Inference Speed.** We evaluated inference speed by comparing RCDMs with AR-LDM and Story-LDM , both diffusion-based methods. All experiments were conducted on the same V100 GPU for consistency. As shown in

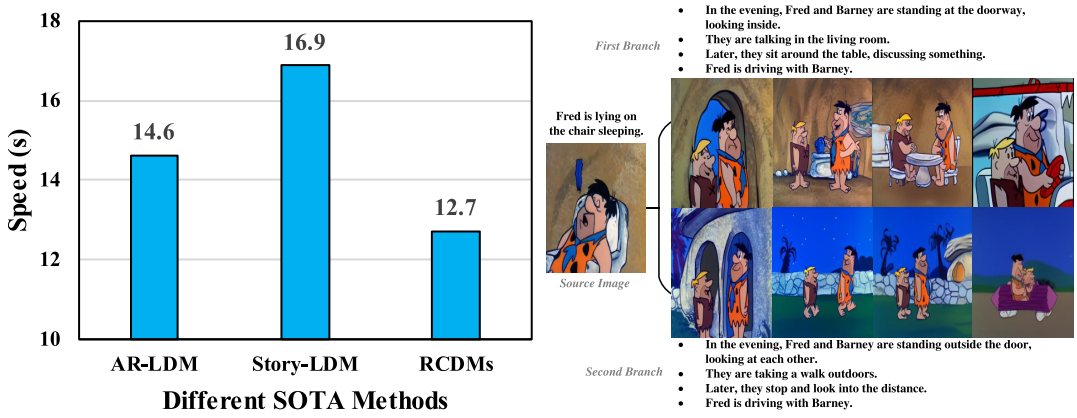


Figure 6: **Left**: Generating stories from different captions in a branching storyline. **Right**: Inference speed comparison with different SOTA methods.

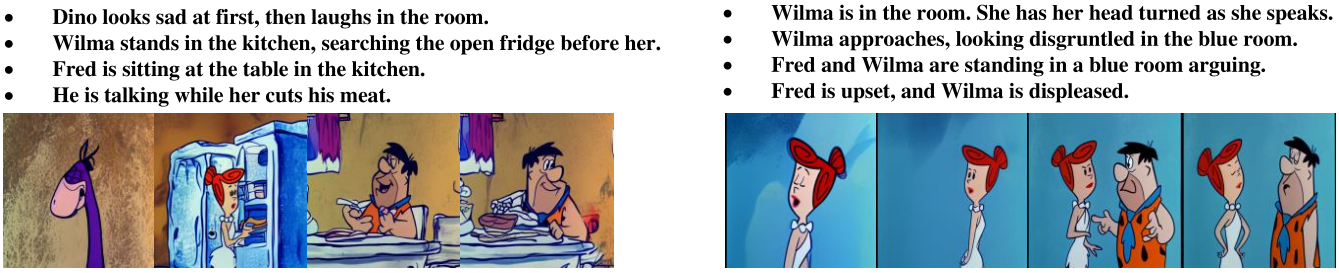


Figure 7: Qualitative results of RCDMs for caption-only story generation.

Figure 6, AR-LDM requires an average of 14.6 seconds per story, and Story-LDM 16.9 seconds, while the two-stage RCDMs complete inference in just 12.7 seconds. The slower performance of AR-LDM and Story-LDM can be attributed to their heavier multimodal models and memory storage mechanisms. Moreover, unlike their autoregressive approaches that generate frames sequentially, RCDMs produce all frames in a single forward pass. These results highlight the architectural efficiency of RCDMs.

**Caption-Only Generation.** RCDMs also support caption-only generation, as it adopts a strategy of randomly dropping images during the training process. In contrast, other SOTA methods typically require an additional model to be trained to accommodate this scenario. Figure 7 displays the results generated by RCDMs using only captions as guidance. Due to the lack of open-source weights from SOTA methods for comparison, we can only assess RCDMs' performance independently. The results suggest that RCDMs can generate story images that maintain consistency in style and sequence under different scenarios/characters.

## Conclusion

We have introduced RCDMs, which leverage rich contextual conditions and diffusion-based modeling to enhance style and temporal consistency in story visualization. Our two-stage approach first aligns semantic correlations between known and unknown clips using a frame-prior transformer diffusion model, then integrates reference frames, predicted semantic embeddings, and text embeddings to achieve coherent, stylistically consistent sequences. Both qualitative and quantitative evaluations show that RCDMs excel in challenging scenarios. **Limitations.** Current methods, including RCDMs, typically achieve consistent story generation on a closed-set dataset, which limits the variety of characters and scenes. For future work, we wil explore methods with open-set generation capabilities to allow for a broader range of characters and scenes.

## Acknowledgments

This work is supported by the Tencent rhino bird elite talent program and the China Scholarship Council program.

## References

Ahn, D.; Kim, D.; Song, G.; Kim, S. H.; Lee, H.; Kang, D.; and Choi, J. 2023. Story visualization by online text augmentation with context memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3125–3135.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

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, Y.; Huang, W.; Zhou, S.; Chen, Q.; and Xiong, Z. 2023. Self-supervised neuron segmentation with multi-agent reinforcement learning. In *IJCAI*, 609–617.

Chen, Y.; Liu, C.; Liu, X.; Arcucci, R.; and Xiong, Z. 2024. Bimcv-r: A landmark dataset for 3d ct text-image retrieval. In *MICCAI*, 124–134. Springer.

Creswell, A.; White, T.; Dumoulin, V.; Arulkumaran, K.; Sengupta, B.; and Bharath, A. A. 2018. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1): 53–65.

Ding, X.; Wang, Y.; Xu, Z.; Welch, W. J.; and Wang, Z. J. 2020. Ccgan: Continuous conditional generative adversarial networks for image generation. In *International conference on learning representations*.

Duan, Y.; Zhao, J.; Mao, J.; Wu, H.; Xu, J.; Ma, C.; Wang, K.; Wang, K.; Li, X.; et al. 2024. Causal Deciphering and Inpainting in Spatio-Temporal Dynamics via Diffusion Model. *arXiv preprint arXiv:2409.19608*.

Ge, S.; Park, T.; Zhu, J.-Y.; and Huang, J.-B. 2023. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7545–7556.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, F.-T.; Zhang, L.; Shen, L.; and Xu, D. 2022. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3397–3406.

Kingma, D. P.; Welling, M.; et al. 2019. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4): 307–392.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Li, B. 2022. Word-level fine-grained story visualization. In *European Conference on Computer Vision*, 347–362. Springer.

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.

Li, Y.; Long, Q.; Zhou, Y.; Cao, N.; Liu, S.; Zheng, F.; Zhu, Z.; Ning, Z.; Xiao, M.; Wang, X.; et al. 2024a. COMAE: COMprehensive Attribute Exploration for Zero-shot Hashing. *arXiv preprint arXiv:2402.16424*.

Li, Y.; Lu, Y.; Dong, Z.; Yang, C.; Chen, Y.; and Gou, J. 2024b. SGLP: A Similarity Guided Fast Layer Partition Pruning for Compressing Large Deep Models. *arXiv preprint arXiv:2410.14720*.

Li, Y.; Wang, H.; Jin, Q.; Hu, J.; Chemerys, P.; Fu, Y.; Wang, Y.; Tulyakov, S.; and Ren, J. 2024c. Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems*, 36.

Li, Y.; Xiao, J.; Feng, C.; Wang, X.; and Chua, T.-S. 2023a. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13869–13878.

Li, Y.; Yang, X.; Zhang, A.; Feng, C.; Wang, X.; and Chua, T.-S. 2023b. Redundancy-aware transformer for video question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3172–3180.

Li, Y.; Zhao, N.; Xiao, J.; Feng, C.; Wang, X.; and Chua, T.-s. 2024d. LASO: Language-guided Affordance Segmentation on 3D Object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14251–14260.

Long, X.; Zeng, J.; Meng, F.; Ma, Z.; Zhang, K.; Zhou, B.; and Zhou, J. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18733–18741.

Luo, J.; Wang, Y.; Gu, Z.; Qiu, Y.; Yao, S.; Wang, F.; Xu, C.; Zhang, W.; Wang, D.; and Cui, Z. 2024. MMM-RS: A Multi-modal, Multi-GSD, Multi-scene Remote Sensing Dataset and Benchmark for Text-to-Image Generation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Maharana, A.; and Bansal, M. 2021. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*.

Maharana, A.; Hannan, D.; and Bansal, M. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

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.

Pan, X.; Qin, P.; Li, Y.; Xue, H.; and Chen, W. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2920–2930.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1505–1514.

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.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

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

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

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans,

T.; et al. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Wei, Y. 2023. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*.

Shen, X.; and Elhoseiny, M. 2023. Large Language Models as Consistent Story Visualizers. *arXiv preprint arXiv:2312.02252*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Wang, C.; Tian, K.; Guan, Y.; Zhang, J.; Jiang, Z.; Shen, F.; Han, X.; Gu, Q.; and Yang, W. 2024a. Ensembling Diffusion Models via Adaptive Feature Aggregation. *arXiv preprint arXiv:2405.17082*.

Wang, C.; Tian, K.; Zhang, J.; Guan, Y.; Luo, F.; Shen, F.; Jiang, Z.; Gu, Q.; Han, X.; and Yang, W. 2024b. V-Express: Conditional Dropout for Progressive Training of Portrait Video Generation. *arXiv preprint arXiv:2406.02511*.

Wang, Y.; Cui, Z.; and Li, Y. 2023. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22025–22034.

Wang, Y.; Li, Y.; and Cui, Z. 2024. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36.

Wu, Q.; Liu, Y.; Zhao, H.; Bui, T.; Lin, Z.; Zhang, Y.; and Chang, S. 2023. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7766–7776.

Xiao, Y.; Wu, L.; Guo, J.; Li, J.; Zhang, M.; Qin, T.; and Liu, T.-y. 2023. A survey on non-autoregressive generation for neural machine translation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xie, J.; Cai, Y.; Chen, J.; Xu, R.; Wang, J.; and Li, Q. 2024a. Knowledge-Augmented Visual Question Answering With Natural Language Explanation. *IEEE Transactions on Image Processing*.

Xie, J.; Chen, J.; Liu, Z.; Cai, Y.; Huang, Q.; and Li, Q. 2024b. Video Question Generation for Dynamic Changes. *IEEE Transactions on Circuits and Systems for Video Technology*.

Xie, J.; Zhou, Z.; Wu, Z.; Zhang, X.; Wang, J.; Cai, Y.; and Li, Q. 2024c. Automated Defect Report Generation for Enhanced Industrial Quality Control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19306–19314.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1316–1324.

Yang, Z.; Wang, J.; Gan, Z.; Li, L.; Lin, K.; Wu, C.; Duan, N.; Liu, Z.; Liu, C.; Zeng, M.; et al. 2023. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14246–14255.

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, J.; Li, Z.; Jiang, H.; and Wang, Z. 2024a. SD-MVS: Segmentation-Driven Deformation Multi-View Stereo with Spherical Refinement and EM Optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6871–6880.

Yuan, Z.; Cao, J.; Wang, Z.; and Li, Z. 2024b. Tsar-mvs: Textureless-aware segmentation and correlative refinement guided multi-view stereo. *Pattern Recognition*, 154: 110565.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-attention generative adversarial networks. In *International conference on machine learning*, 7354–7363. PMLR.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 833–842.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 5907–5915.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.

Zhang, Z.; Sun, S.; Wang, W.; Cai, D.; and Bian, J. 2024. FlexCAD: Unified and Versatile Controllable CAD Generation with Fine-tuned Large Language Models. *arXiv preprint arXiv:2411.05823*.

Zhou, Y.; Li, Y.; Feng, L.; and Huang, S.-J. 2024. Improving Generalization of Deep Neural Networks by Optimum Shifting. *arXiv preprint arXiv:2405.14111*.