Story Augmentation with Generative AI (SAGANets): Investigating Multi-Image Story-Generation Pipelines

Connor Janowiak
Department of Computer Science
Stanford University
cjanow@stanford.edu

Bradley Moon
Department of Computer Science
Stanford University
bradmoon@stanford.edu

Sadé Ried
Department of Computer Science
Stanford University
saderied@stanford.edu

Abstract

Recent advances in generative models have shown remarkable success in synthesizing images from text. However, generating a sequence of images that coherently illustrate a story remains a significant challenge, primarily due to the need for visual and narrative consistency across multiple scenes. This project introduces SAGANets, a multimodal pipeline designed to explore automated story generation. SAGANets alternates between generating a caption using a fine-tuned GPT-2 model and synthesizing the corresponding image via a Stable Diffusion 1.5 UNet with LoRA adapters. We trained and evaluated this pipeline on the FlintstonesSV dataset, aiming to produce multi-frame mini-stories conditioned on a variable-length prefix of caption-image pairs. Our experiments with SAGANets revealed challenges, including rapid overfitting of the LoRA-adapted image model and difficulties in maintaining robust narrative progression with the captioning module, often resulting in abstract or inconsistent visuals. In parallel, we explored alternative approaches, notably a GAN-based model, which demonstrated more promising results in terms of visual fidelity and character consistency on this dataset. This paper presents the SAGANets architecture, discusses its performance, and contrasts it with these alternatives, highlighting key learnings and future directions for automated multi-image storytelling.

1. Introduction

The emergence of deep generative models, especially Generative Adversarial Networks (GANs) and more recently, Denoising Diffusion Probabilistic Models (DDPMs), has significantly advanced the field of artificial intelligence, particularly in image generation. Prominent models like DALL-E and Stable Diffusion [6] have demonstrated an impressive capability to generate high-fidelity images from textual descriptions. These innovations have spurred numerous applications, from artistic creation and design to educational tools and content generation.

Despite these advancements, much of the current research focuses on generating a single image from a textual prompt. The more complex task of creating a sequence of images to visually narrate a story presents substantial difficulties. These primarily arise from the critical need to maintain visual consistency of characters, objects, and settings across multiple frames, ensure each image accurately portrays its narrative segment, and guarantee a coherent flow in the sequence. As Kim [2] notes, existing methods often struggle with these interconnected challenges.

Figure 1. Example illustrating the challenge of maintaining consistency in story visualization, here generated by a GAN-based approach explored in our study.

In this project, we develop and investigate SAGANets, a story-generation pipeline that iteratively generates captions using GPT-2 and corresponding images using Stable Diffusion 1.5 with LoRA adapters. Our goal was to produce coherent multi-frame stories on the FlintstonesSV dataset from an initial prompt. This paper details the SAGANets architecture, analyzes its performance, discusses the encoun-

tered challenges, and compares its outcomes with alternative methodologies explored during our research, offering insights into the complexities of automated story generation.

2. Related Work

Early efforts in story visualization include StoryGAN [3], which framed the problem as sequential image generation conditioned on a multi-sentence story. It featured a recurrent Context Encoder and dual GAN discriminators (image and story level) to enhance local fidelity and global narrative consistency. While notable for its time, StoryGAN's GAN-based image generation often resulted in lower image quality compared to contemporary diffusion models and faced challenges in aligning complex narrative elements with visuals.

More recently, StoryDALL-E [4] proposed adapting large pretrained text-to-image transformers for story continuation. This involves generating an image sequence based on captions and a source frame, retrofitting models with custom modules for sequential generation and cross-attention layers to copy visual elements, thereby improving continuity. Their work highlighted that pretrained transformers can be fine-tuned for multi-image generation, outperforming GAN baselines like StoryGAN.

Our work on SAGANets draws inspiration from such prior efforts but is distinguished by its specific pipeline: we aim to coordinate a fine-tuned GPT-2 for next-caption prediction with a Stable Diffusion 1.5 UNet fine-tuned using LoRA adapters for image generation, alternating between these modules. Unlike StoryDALL-E, which often relies on an explicit source frame at inference, SAGANets is designed to condition generation on a variable-length prefix. Furthermore, by operating in Stable Diffusion's latent space and fine-tuning only lightweight LoRA modules and GPT-2, we aimed for more efficient training, particularly relevant under limited computational resources.

3. Data

In this project, we use the `FlintstonesSV` dataset [3], comprising brief, densely annotated clips detailing actions, characters, objects, and settings. Annotations include character identification and localization, scene setting, captions, object annotation, and entity tracking across frames. The dataset provides both visual and textual components.

The `flintstones_annotations_v1.json` file contains captions for each clip. The `flintstones_dataset.tgz` files contain video clips, background frames, entity segmentation, tracking information, and video frames as numpy arrays (`num_frames`, $H, W, 3$). Each story consists of caption-image pairs. A story instance S_j is:

$$S_j = ((I_{j,1}, \ldots, I_{j,N_{\text{frames}}}), (C_{j,1}, \ldots, C_{j,N_{\text{frames}}}))$$

where $I_{j,n} \in \mathbb{R}^{H \times W \times 3}$ is the n-th image, $C_{j,n}$ is its caption, and $N_{\text{frames}} = N$.

We created a custom `StoryImageDataset` to structure this data for our training pipeline. Each sample contains the first frame, a list of subsequent frames (images 2-N), tokens for subsequent captions, and sentence embeddings. This is reassembled as `N_frames` $=$ [`first_frame`] $+$ `future_frames` and `N_captions` $=$ [`caption1`] $+$ `tokens` for processing by our models.

4. Methods

Our primary investigated pipeline, SAGANets, consists of two main components: a caption generation module and an image generation module, which are intended to operate sequentially.

4.1. GPT-2 Caption Module

The captioning module aims to predict the $(k+1)$-th caption given a serialized prefix of prior captions. The input format is:

$$P_k = [S_1, C_1, S_2, C_2, \ldots, S_k, C_k, S_{\text{NXT}}]$$

where S_i is $\langle \text{SCENE} \rangle_i$ and C_j is $\langle \text{CAPTION} \rangle_j$. The model autoregressively generates tokens w_j for $C_{k+1} = (w_1, \ldots, w_m)$ using cross-entropy loss, applied only to target caption tokens:

$$\mathcal{L}_{\text{caption}} = -\sum_{j=1}^{m} \log p(w_j^* \mid P_k, w_1^*, \ldots, w_{j-1}^*; \theta_{\text{GPT}})$$

where $C_{k+1}^* = (w_1^*, \ldots, w_m^*)$ is the ground truth caption, and $p(\cdot)$ is the probability predicted by the GPT-2 model (parameters θ_{GPT}). We used a pretrained GPT-2, fine-tuned on the FlintstonesSV dataset.

4.2. Stable Diffusion 1.5 UNet & LoRA Module

For generating frame I_{k+1}, we use a latent diffusion model. An initial latent z_0 is noised over T timesteps to z_T. The model predicts the noise ε added to get z_t at timestep t, conditioned on z_t and a vector c_{k+1} from caption$_{k+1}$. This predicted noise is $\varepsilon_{\theta_{SD}}(z_t, t, c_{k+1})$. During generation, a random latent is iteratively denoised to z_0, then decoded to image I_{k+1}.

Our backbone is the UNet from `runwayml/stable-diffusion-v1-5`. Base UNet weights W_0 are frozen. We integrate Low-Rank Adaptation (LoRA) [1] into attention layers (W_q, W_k, W_v). LoRA approximates weight updates ΔW using $W' = W_0 + BA$, where B and A are smaller, trainable matrices. For instance, $W_q' = W_q + B_q A_q$. Captions C_{k+1} are embedded using Stable Diffusion's CLIP text encoder $\mathcal{E}_{\text{CLIP}}$ to get $c_{k+1} \in \mathbb{R}^{768}$. The training objective is the MSE loss:

$$\mathcal{L}_{\text{image}} = \mathbb{E}_{z_0, c_{k+1}, \varepsilon \sim \mathcal{N}(0,I), t} \|\varepsilon - \varepsilon_{\theta_{SD}}(z_t, t, c_{k+1})\|_2^2$$

Here, $z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon$. θ_{SD} includes trainable LoRA adapters θ_{LoRA}. The VAE from Stable Diffusion 1.5 handles image encoding/decoding.

4.3. Coordinated Training of SAGANets Modules

The GPT-2 caption module and the Stable Diffusion LoRA-adapted image module are fine-tuned based on the FlintstonesSV dataset. During training, for a given story, we sample a random prefix length $k \sim \mathcal{U}(\{0, 1, 2, 3\})$. The GPT-2 module is trained to predict the ground truth caption C_{k+1} given the ground truth prior captions $C_1, \ldots, C_k$. The Stable Diffusion UNet (with LoRA adapters) is trained to generate the ground truth image I_{k+1} conditioned on the ground truth caption C_{k+1} (embedded as c_{k+1}). This component-wise training with ground truth inputs means that errors from a generated caption are not propagated to the image module during this training phase. Only the GPT-2 parameters and the LoRA adapters are updated. This strategy was chosen for training stability and to leverage the distinct strengths of pretrained models for their respective tasks. The "alternating generation" is thus primarily an inference-time procedure.

4.4. Alternative Approaches Explored

In addition to the SAGANets pipeline, we investigated several alternative methodologies to provide a broader context for the challenges in sequential image generation. These approaches, common in generative modeling, offer different philosophies for image synthesis and consistency.

GAN-Based Approach: This framework employs a generator network G (in our case, a U-Net architecture derived from Stable Diffusion and fine-tuned with LoRA) and a dedicated Convolutional Neural Network (CNN) as a discriminator D. The core idea is an adversarial process: D is trained to distinguish real images x from the dataset versus synthetic images $G(z)$ produced by the generator from a latent vector z. Concurrently, G is trained to produce images that D misclassifies as real. This min-max game is defined by:

$$\min_G \max_D \mathcal{V}(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

Beyond the adversarial loss $\mathcal{L}_{\text{adv-G}}$, which encourages realism, our generator's total loss function $\mathcal{L}_G = \mathcal{L}_{\text{adv-G}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{char}}\mathcal{L}_{\text{char}}$ incorporated auxiliary losses. $\mathcal{L}_{\text{img}}$ (e.g., an L_1 loss) provides direct pixel-level guidance, stabilizing training and ensuring content preservation. $\mathcal{L}_{\text{clip}}$ leverages CLIP's joint text-image embedding space to enforce semantic consistency between the generated image and its target caption. $\mathcal{L}_{\text{char}}$ aims to preserve character appearance across frames using a feature extractor, crucial for narrative continuity. This multifaceted loss helps balance image sharpness and realism, learned implicitly by the discriminator, with explicit content and semantic guidance.

CLIP-Centric Approach: Motivated by the power of CLIP [5] in aligning images and text, this approach forgoes adversarial training and focuses on direct optimization of image content based on semantic and perceptual losses. The generator G (e.g., a diffusion model like Stable Diffusion or a GAN generator backbone) is optimized using a composite loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{img}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{char}}\mathcal{L}_{\text{char}}$$

Here, $\mathcal{L}_{\text{img}}$ is typically an L_2 image reconstruction loss encouraging fidelity to a target image if available, or promoting general image quality. The core $\mathcal{L}_{\text{clip}}$ term, $1 - \texttt{cos\_sim}(\mathcal{E}_{\text{img}}(G(z)), \mathcal{E}_{\text{text}}(\text{caption}))$, directly steers the image generation towards the semantic meaning of the caption as interpreted by CLIP's image ($\mathcal{E}_{\text{img}}$) and text ($\mathcal{E}_{\text{text}}$) encoders. The character consistency loss $\mathcal{L}_{\text{char}}$ can be formulated to ensure that character features, extracted by f_{char}, remain similar across a sequence: $\sum_{k=1}^{N-1} \| f_{\text{char}}(G(z)_k) - f_{\text{char}}(G(z)_{k+1}) \|_2$. This method prioritizes strong text-image correspondence and character persistence, leveraging CLIP's rich learned representations, but can sometimes lead to artifacts if CLIP's understanding is "exploited" by the generator, or if loss components are not well-balanced.

Temporal Coherence Approach: This method was specifically conceptualized to enforce smooth visual transitions and continuity between consecutive frames, a key aspect of believable story sequences. The emphasis is on maintaining consistency in a learned latent space. Given a sequence of generated frames $G(z_k)$, the ir latent representations $L_k = E(G(z_k))$, obtained via an encoder E (which could be part of a VAE or a separate pretrained network), are constrained. The total loss for this approach might be:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{img}} + \lambda_{\text{coher}}\mathcal{L}_{\text{coher}}$$

$\mathcal{L}_{\text{img}}$ is again an L_2 image reconstruction loss. The distinctive component is the temporal coherence loss $\mathcal{L}_{\text{coher}}$:

$$\mathcal{L}_{\text{coher}} = \sum_{k=1}^{N-1} \| L_{k+1} - L_k \|_2^2$$

This loss explicitly minimizes the Euclidean distance between latent representations of adjacent frames, promoting gradual changes in visual features. The choice of encoder E and its latent space is crucial, as different spaces capture different levels of abstraction. While theoretically promising for enhancing sequence smoothness, our attempts to implement this encountered technical challenges during development, particularly with debugging the inference script for

sequential generation under this loss. These implementation hurdles unfortunately prevented us from fully realizing and evaluating a functional version within the project's timeframe.

Finally, a *Combined Strategy* was also envisioned, which would integrate elements from all three: GAN-based training for image sharpness and quality, CLIP-based losses for robust semantic and character consistency, and temporal coherence losses for frame-to-frame smoothness. Such a model would optimize a comprehensive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv-G}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}} + \lambda_{\text{char}}\mathcal{L}_{\text{char}} + \lambda_{\text{coher}}\mathcal{L}_{\text{coher}}$$

This multifaceted approach aims to balance diverse objectives, but its complexity also implies significant challenges in tuning the various hyperparameters λ to achieve synergy rather than conflicting gradients. This was not implemented due to time constraints. These alternatives were explored to benchmark different architectural philosophies for story generation against our primary SAGANets pipeline.

5. Experiments and Results

We conducted experiments to evaluate the SAGANets pipeline and the alternative approaches on the FlintstonesSV dataset.

5.1. SAGANets Performance

Our initial experiments training the SAGANets pipeline (joint GPT-2 & LoRA-SD-1.5 model) yielded mixed results. After training for one epoch, the model could produce images resembling abstract cartoon scenes (Figure 2). However, these generations often lacked specific character details from the Flintstones universe, like Fred, Wilma, or Barney, depicting more generic cartoon figures.

Figure 2. Example output from SAGANets after one epoch of training, showing three generated scenes. The style is cartoonish but character specificity is limited.

We observed that using a smaller LoRA scaling factor (α) at inference than during training ($\alpha = 32$) improved output quality. Figure 3 shows that a LoRA-$\alpha = 10$ at inference produced more comic-like images compared to $\alpha = 32$, suggesting the lower α helped mitigate overfitting and improve visual coherence. The LoRA update is $\Delta W = \frac{\alpha}{r}BA$, so $W' = W_0 + \frac{\alpha}{r}BA$.

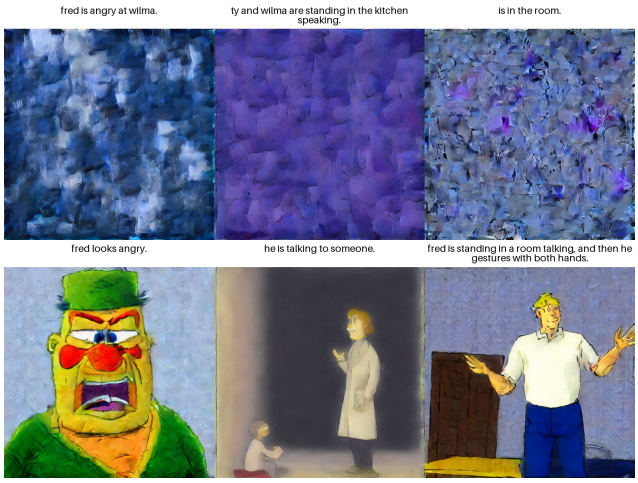

Figure 3. Comparison of SAGANets outputs using different LoRA-α values at inference. Top: $\alpha = 32$ (matching training). Bottom: $\alpha = 10$. The lower α yields visually preferable results.

However, extending training beyond one epoch for the LoRA-adapted UNet led to significant degradation. As shown in Figure 4, outputs became extremely noisy and unrecognizable, indicating rapid overfitting despite the image-generation loss not collapsing.

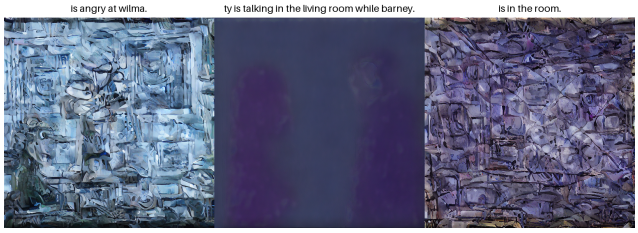

Figure 4. Example of SAGANets output when trained for more than one epoch, showing noisy and unrecognizable results due to overfitting.

The fine-tuned GPT-2 captioning module also presented limitations. While it could generate contextually related captions, its ability to creatively and coherently extend storylines was inconsistent. We observed instances of incomplete or non-sensical captions, particularly when trying to meet a target sequence length (Figure 5). This poor caption quality likely compounded the issues faced by the sensitive, overfit image model.

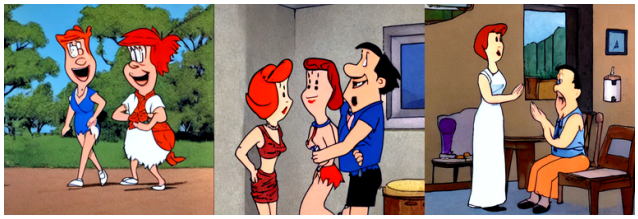

Figure 5. Example demonstrating challenges in GPT-2 caption generation and subsequent image synthesis coherence within the SAGANets pipeline after a single epoch.

5.2. Performance of Alternative Approaches

GAN-Based Approach: This approach yielded the most promising results among the alternatives in terms of visual output quality and training stability. The generator and discriminator losses achieved a stable balance (approximate mean $g_{loss} \approx 1.197$, $d_{loss} \approx 0.686$ after convergence, see Figure 6). Generated images exhibited commendable visual quality and, importantly, showed better capability in maintaining character consistency across frames (see Figure 7).

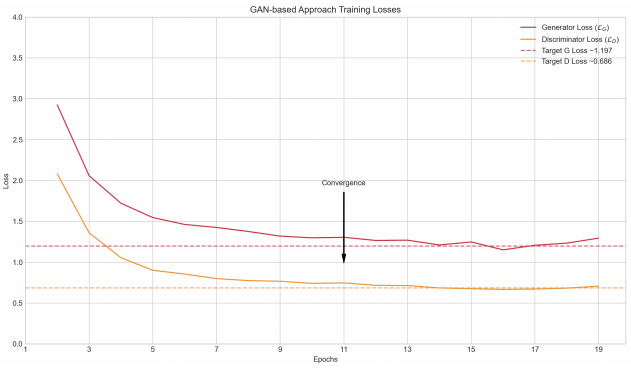

Figure 6. Training curves for the GAN-based model, showing generator and discriminator losses over epochs.

Figure 7. Image generation comparison from the GAN-based alternative. Top: 20 training epochs. Bottom: 5 training epochs. Shows reasonable character consistency.

CLIP-Centric Approach: This approach, while initially promising for semantic alignment, quickly became unstable. The CLIP loss oscillated and failed to converge (Figure 8), correlating with degraded image quality suffering from artifacts, despite some semantic relevance. Due to persistent overfitting and inability to achieve competitive visual quality, this line was discontinued.

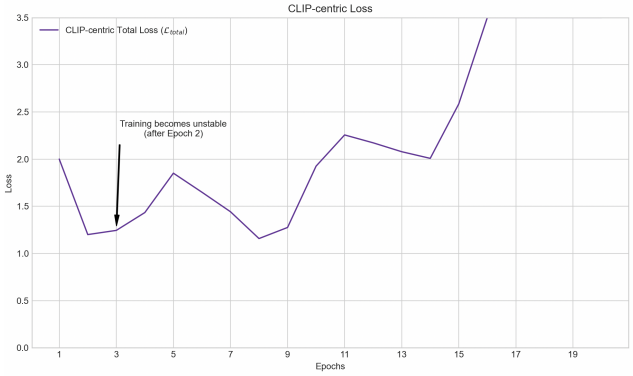

Figure 8. Training curve for the CLIP-centric model, illustrating the loss trend before training instability.

Temporal Coherence Approach: As noted in Section 4.4, implementation challenges prevented a full evaluation of this method.

5.3. Discussion of Experimental Findings

The experiments with SAGANets highlighted significant challenges in fine-tuning large pretrained models for sequential, conditioned generation on specialized datasets like FlintstonesSV using LoRA. Rapid overfitting in the image module and coherence issues in the captioning module limited its effectiveness. In contrast, the GAN-based alternative, while computationally more intensive in its traditional setup, demonstrated better robustness in generating visually consistent and higher-fidelity images for this particular task. This suggests that adversarial training might offer advantages in enforcing perceptual quality and consistency in this specific story generation context, or that the chosen LoRA configuration and training strategy for SAGANets were not optimal for FlintstonesSV.

6. Conclusion

In this project, we designed and investigated SAGANets, a multi-image story generation pipeline employing GPT-2 for captioning and a LoRA-adapted Stable Diffusion 1.5 model for image synthesis. Our aim was to generate coherent, multi-frame stories. The evaluation on the FlintstonesSV dataset revealed that while SAGANets could produce stylized cartoon outputs, it faced considerable challenges, including rapid overfitting of the image generation module after minimal training, and limitations in the captioning module's ability to maintain robust narrative progression and coherence. These issues resulted in outputs that were often abstract and lacked consistent character depiction.

Our exploration of alternative approaches provided valuable comparative insights. Notably, a GAN-based model demonstrated superior performance in terms of visual quality, stability during training, and character consistency across frames on this dataset. A CLIP-centric approach,

while conceptually targeting strong semantic alignment, suffered from training instability and failed to produce competitive results.

The findings from SAGANets highlight the complexities of fine-tuning large pretrained diffusion models with methods like LoRA for sequential, conditional tasks. The sensitivity to overfitting and the difficulty in achieving deep narrative and visual coherence with the chosen SAGANets architecture show that there is much work to be done along these lines. While the promise of leveraging powerful pretrained models remains, strategies to improve robustness, control overfitting (perhaps through different LoRA configurations or regularization techniques), and enhance the synergy between text and image generation modules are definitely open paths to continue this work.

Future work could focus on several directions:

(i) Investigating more advanced fine-tuning techniques or regularization methods for the diffusion model component of SAGANets to combat overfitting.

(ii) Exploring alternative architectures for the captioning module or different prompting strategies to improve narrative coherence and relevance.

(iii) Evaluating the impact of dataset size and specificity on the performance of such pipelines.

Overall, we think this work contributes to understanding the practical challenges and comparative strengths of different generative strategies for the nuanced task of automated story visualization, emphasizing that while powerful individual models exist, their effective orchestration into a coherent storytelling pipeline is a difficult task and is definitely an open area of research.

References

[1] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. arXiv:2106.09685. 2

[2] J. Kim, J. Kim, M. Kim, and G. Kim. Story-to-image: A framework for story visualization with character consistency. *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1337–1347, 2023. 1

[3] Y. Li, Z. Gan, Y. Cheng, J. Liu, Y. Shen, L. Wang, J. Zhang, and J. Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6405–6414, 2019. 2

[4] A. Maharana, X. Li, X. Kong, Y. Yang, P. Isola, C. Lu, Y. Wang, and Y. Lu. Improving consistency in multimodal generation: A storytelling perspective. *arXiv preprint arXiv:2209.06192*, 2022. 2

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1