StyleT2V0: Zero-Shot Short Video Generation Guided by Text and Style

Astghik Chobanyan[1,3], Levon Khachatryan[1,2]
[1]Yerevan State University (YSU) [2]Picsart AI Resarch (PAIR) [3]Cognaize Engineering

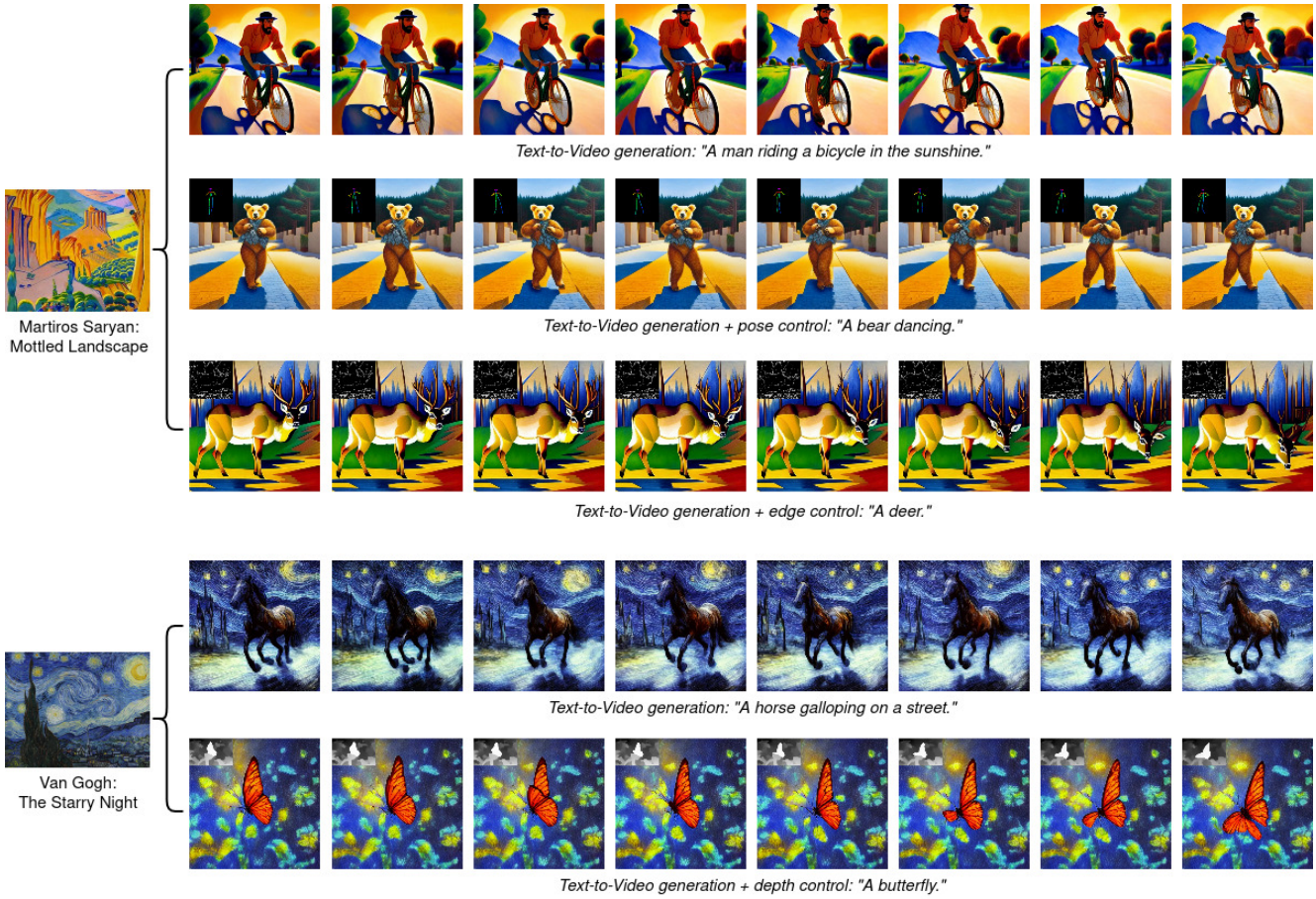

Figure 1. Our method **StyleT2V0** advances video generation capabilities by utilizing: (i) a textual prompt combined with a style image to capture the ensure stylistic and contextual accuracy across frames (see rows 1, 4), (ii) pose control to ensure dynamic accuracy, (iii) edge control for enhanced visual definition, and (iv) depth control to enrich spatial perception. The showcased frames illustrate the method's robust ability to adaptively integrate enhanced controls while consistently reflecting the artistic intent and textual specifications.

Abstract

Recent advancements in text-to-image diffusion models have markedly expanded the potential for generating images based on textual descriptions. Yet, the extension of these capabilities to video generation poses significant challenges, particularly in achieving stylistic consistency across video frames. Traditional approaches often require extensive training processes and face difficulties in maintaining a coherent artistic style throughout the sequence, especially when attempting to emulate a ***specific style referenced by an image****.*

To bridge this gap, we introduce StyleT2V0, a zero-shot methodology that enables the generation of videos directly

from textual prompts while ensuring stylistic alignment with a designated reference image. By integrating shared attention mechanisms within the generation process, our approach guarantees that each frame not only adheres contextually to the textual description but also maintains stylistic coherence with the reference image, all without the need for additional model training. This method represents a significant step forward in simplifying the video production process from textual descriptions, offering a streamlined and efficient solution for creating stylistically consistent video content.

1. Introduction

The advent of generative artificial intelligence (AI) has marked a new era of digital content creation, enabling the synthesis of complex and creative outputs from simple inputs. However, the field of text-to-video generation, despite its potential, grapples with significant challenges due to the inherently complex nature of videos and the intricate relationship between text and visual content. Ensuring temporal and contextual consistency in video generation remains a major hurdle, as videos require each frame to be visually coherent with adjacent frames and aligned with the text prompt provided. Furthermore, maintaining style consistency across all frames introduces an additional layer of complexity.

Current text-to-video generation systems often struggle to capture the desired style effectively. This is partly because conveying style through textual descriptions lacks the explicitness and spatial specificity provided by visual references. The StyleCrafter [16] introduces several innovative strategies to overcome existing limitations in video generation. It leverages a style control adapter, trained with image datasets, to infuse style features from a reference image into video content. It incorporates a Scale-Adaptive Fusion Module, which balances text-based content features with image-based style features. Despite its advanced capabilities, StyleCrafter requires extensive training. It uses a style control adapter that must be trained with image datasets to effectively transfer style features from reference images into the generated video. The method also involves fine-tuning of the T2V models to improve their temporal dynamics and ensure consistent style throughout the video, enhancing the overall quality.

Other techniques in the field utilize various forms of style transfer technology, embedding style features directly within the generative process. These methods typically employ deep learning architectures that integrate style information at different layers of the network, allowing for the dynamic adjustment of style elements during the video generation process. However, these methods can suffer from high computational costs and may require significant tuning to balance style preservation with the visual output.

Despite these challenges, there are promising opportunities for advancements in text-to-video generation technology. Leveraging zero-shot learning could potentially enable models to generate content without prior specific training on similar tasks, thus enhancing the model's flexibility and applicability. Additionally, improving methods for consistent style information throughout the video could lead to more cohesive and aesthetically pleasing outputs. In response to these challenges, we introduce StyleT2V0, a zero-shot and versatile approach that facilitates the creation of videos in any desired style using a reference image (see Fig. 1 and further results). This method offers two key benefits: it enhances the styling capabilities of text-to-video (T2V) models without prior training and ensures a more precise representation of the intended style than text descriptions alone, leading to videos that are not only consistent but also uniform in stylistic execution.

The experiments demonstrate that StyleT2V0 successfully generates temporally consistent short videos from text while ensuring stylistic alignment with a specified reference image. To summarize, our contributions are three-fold:

- A novel attention sharing mechanism to simultaneously enforce style guidance and ensure frame consistency during generation in a zero-shot manner.
- A range of applications demonstrating the effectiveness of our approach, such as generating videos conditioned on a style image alongside edge, pose, or depth maps.
- Our approach demonstrates superior performance in terms of color transformation from style image while preserving the text alignment.

2. Related Work

2.1. Text-to-Image Generation

The field of text-to-image synthesis has witnessed transformative advancements with the development of sophisticated models that generate highly realistic and semantically accurate images directly from textual descriptions. Generative Adversarial Networks (GANs) [6] have been pivotal in pioneering early developments in text-to-image generation. Innovations such as StackGAN [33] and AttnGAN [32] have laid foundational work by using stacked generative models and attention mechanisms, respectively, to progressively refine images from coarse-to-fine details based on textual inputs. These models excelled in generating visually appealing images that closely align with the given text descriptions, setting a new benchmark for image quality and text alignment in early stages. The adoption of transformers, initially designed for text processing, into image synthesis marked a significant leap. Models like DALL-E [20] from OpenAI leveraged transformers to handle complex, abstract concepts and create images with unprecedented creativity

from textual prompts. The model's ability to parse and interpret nuanced text has enabled it to generate images that are not only high in quality but also rich in context and imagination. Diffusion models have recently revolutionized the text-to-image landscape by offering another layer of sophistication. Unlike GANs, diffusion models convert noise into structured images gradually, controlled by the conditioning provided by text. This process has proven to produce images with exceptional detail and lower rates of artifacts. GLIDE [18] and Imagen [25] are notable examples that use diffusion techniques to achieve remarkable levels of realism and accuracy, outperforming previous methodologies in various benchmarks. Advancing further, Latent Diffusion Models (LDMs) [22] operate by encoding images in a compressed latent space before applying the diffusion process, significantly enhancing computational efficiency. This method maintains the high quality of generated images while reducing the resource intensity typically associated with such processes. The use of latent space allows these models to manage and manipulate higher-resolution images more effectively, facilitating broader applications.

Advancing further, Latent Diffusion Models (LDMs) [22] operate by encoding images in a compressed latent space before applying the diffusion process, significantly enhancing computational efficiency. This method maintains the high quality of generated images while reducing the resource intensity typically associated with such processes. The use of latent space allows these models to manage and manipulate higher-resolution images more effectively, facilitating broader applications. Building upon the capabilities of LDMs, methodologies such as Diffusion-Enhanced PatchMatch [7] and StyleAlign [9] have emerged, demonstrating the adaptability of diffusion models in style transfer. Diffusion-Enhanced PatchMatch leverages the model's diffusion properties to ensure coherent style application across images through targeted style patch matching and blending. In contrast, StyleAlign employs minimal attention-sharing mechanisms throughout the diffusion process to produce style-consistent sets of images, proving effective in maintaining stylistic consistency without extensive training or fine-tuning. These advancements underscore the ongoing evolution and increasing complexity of generative models in the text-to-image synthesis arena.

2.2. Text-to-Video Generation

Text-to-video generation continues to be a vibrant and challenging area of research within artificial intelligence, aiming to transform textual descriptions into dynamic visual narratives. This field has evolved significantly, initially leveraging generative adversarial networks (GANs) [6] and vector quantized variational autoencoders (VQ-VAE) [28] to more advanced methodologies employing transformer-based architectures and diffusion models.

The utilization of autoregressive transformers has profoundly impacted the development of this technology. Models such as NUWA [31] and Phenaki [29] are at the forefront, with NUWA implementing a 3D transformer encoder-decoder framework that supports both image and video generation from text. Phenaki advances this paradigm using a bidirectional masked transformer that employs a causal attention mechanism, enabling the synthesis of extended video sequences from concise text descriptions. These advancements underscore the versatility of transformers in handling the complexities of video content generation.

In parallel, diffusion models have brought substantial enhancements to the fidelity and coherence of video outputs. CogVideo [14] extends the principles of the CogView2 [3] model, incorporating structured multi-frame-rate hierarchical training strategies to refine the alignment of text and video. Large-scale diffusion models like those used in Video Diffusion Models (VDM) [13] and Imagen Video [12] apply cascading diffusion models tailored for video, achieving high-resolution outputs that maintain temporal consistency.

Further expanding the model ecosystem, approaches like Make-A-Video [26] and Gen-1 [5] explore unsupervised learning paradigms and content-guided video editing, respectively. Make-A-Video leverages existing text-to-image architectures, applying them to video synthesis in an unsupervised manner. Conversely, Gen-1 introduces a novel framework for video editing that is guided by structural and content-based descriptions, facilitating refined control over the generated video content.

The introduction of diffusion transformers has further revolutionized video generation, leading to solutions like Latte [17] and Sora [1], which can produce minute-long videos of high visual quality that faithfully follow human instructions. These models demonstrate remarkable capabilities in interpreting and visualizing complex scenarios presented in textual form.

Recent innovations have introduced zero-shot, training-free methodologies to this discipline. Text2Video-Zero [15] exemplifies this approach by adapting pre-existing text-to-image models to generate video sequences without extensive retraining, significantly reducing the computational footprint. This model enhances latent codes with motion dynamics and integrates cross-frame attention, ensuring temporal consistency across generated video frames from textual prompts. However, Text2Video-Zero lacks the capability for style image guidance, thus necessitating the training of a DreamBooth [24] model for each unique style to facilitate video generation with the desired stylistic characteristics. The StyleCrafter [16], on the other hand, introduced a style control adapter, trained with image datasets, to infuse style features from a reference image into video content. It incorporated a Scale-Adaptive Fusion Module, which bal-

anced text-based content features with image-based style features. Despite its advanced capabilities, StyleCrafter requires extensive training. It uses a style control adapter that must be trained with image datasets to effectively transfer style features from reference images into the generated video. The method also involves fine-tuning of the T2V models to improve their temporal dynamics and ensure consistent style throughout the video, enhancing the overall quality.

Unlike the aforementioned methods, StyleT2V0 is entirely training-free, eliminating the necessity for significant computing resources or multiple GPUs, all while allowing for style image guidance during the generation process.

3. Preliminaries

3.1. Stable Diffusion (SD)

SD is a diffusion model operating in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, namely VQ-GAN [4] or VQ-VAE [28], where $\mathcal{E}$ and $\mathcal{D}$ are the corresponding encoder and decoder, respectively. More precisely if $x_0 \in \mathbb{R}^{h\times w\times c}$ is the latent tensor of an input image Im to the autoencoder, i.e. $x_0 = \mathcal{E}(Im)$, diffusion forward process iteratively adds Gaussian noise to the signal x_0:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),\ t = 1, .., T \quad (1)$$

where $q(x_t|x_{t-1})$ is the conditional density of x_t given x_{t-1}, and $\{\beta_t\}_{t=1}^T$ are hyperparameters. T is chosen to be as large that the forward process completely destroys the initial signal x_0 resulting in $x_T \sim \mathcal{N}(0, I)$. The goal of SD is then to learn a backward process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

for $t = T, \dots, 1$, which allows to generate a valid signal x_0 from the standard Gaussian noise x_T. To get the final image generated from x_T it remains to pass x_0 to the decoder of the initially chosen autoencoder: $Im = \mathcal{D}(x_0)$.

After learning the above mentioned backward diffusion process (see DDPM [11]) one can apply a deterministic sampling process, called DDIM [27]:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t), \quad t = T, \dots, 1, \quad (3)$$

where $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ and

$$\epsilon_\theta^t(x_t) = \frac{\sqrt{1-\alpha_t}}{\beta_t}x_t + \frac{(1-\beta_t)(1-\alpha_t)}{\beta_t}\mu_\theta(x_t, t). \quad (4)$$

To get a text-to-image synthesis framework, SD guides the diffusion processes with a textual prompt τ. Particularly for DDIM sampling, we get:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t, \tau)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t, \tau), \quad t = T, \dots, 1. \quad (5)$$

It is worth noting that in SD, the function $\epsilon_\theta^t(x_t, \tau)$ is modeled as a neural network with a UNet-like [23] architecture composed of convolutional and (self- and cross-) attentional blocks. x_T is called the latent code of the signal x_0 and there is a method [2] to apply a deterministic forward process to reconstruct the latent code x_T given a signal x_0. This method is known as DDIM inversion. Sometimes for simplicity, we will call $x_t, t = 1, \dots, T$ also the *latent codes* of the initial signal x_0.

3.2. Text2Video-Zero

Our baseline model, which is termed ***Text2Video-Zero***[15], is a state-of-the-art framework designed to generate videos directly from textual descriptions without requiring any fine-tuning or pre-training on video datasets. Text2Video-Zero harnesses the synergy of motion dynamics and cross-frame attention mechanisms to forge a path toward seamless text-to-video generation, ensuring both temporal consistency and fidelity to the input text description.

The process begins with a randomly sampled latent code x_1^T for the initial frame, which is then refined using Δt DDIM [27] backward steps to derive $x_1^{T'}$, utilizing a pre-trained Stable Diffusion model (SD) [21]. To incorporate motion dynamics, a specified motion field is applied across the video sequence, employing a warping function W_k for each frame k, which transforms $x_1^{T'}$ into $x_k^{T'}$. This step is pivotal as it incorporates the latent codes with the requisite motion dynamics, thereby dictating the global scene and camera motion. This ensures temporal consistency across the background and the overall scene, enhancing the clarity of the generated video.

Subsequently, the denoised latent codes x_k^T for each frame k are obtained through a forward application of the DDPM [11] process. This probabilistic method allows for a greater degree of freedom in managing the motion of objects within the video, further contributing to a dynamic and realistic representation of the scene.

The core of Text2Video-Zero's capability to maintain visual continuity and object identity lies in its innovative use of cross-frame attention. By employing keys and values derived from the first frame's latent code, the model generates subsequent frames ($k = 1, \dots, m$) that not only share stylistic and narrative elements but also preserve the appearance and identity of foreground objects across the sequence. This cross-frame attention is crucial for ensuring that each frame contributes to a cohesive video narrative, reflecting the text prompt accurately.

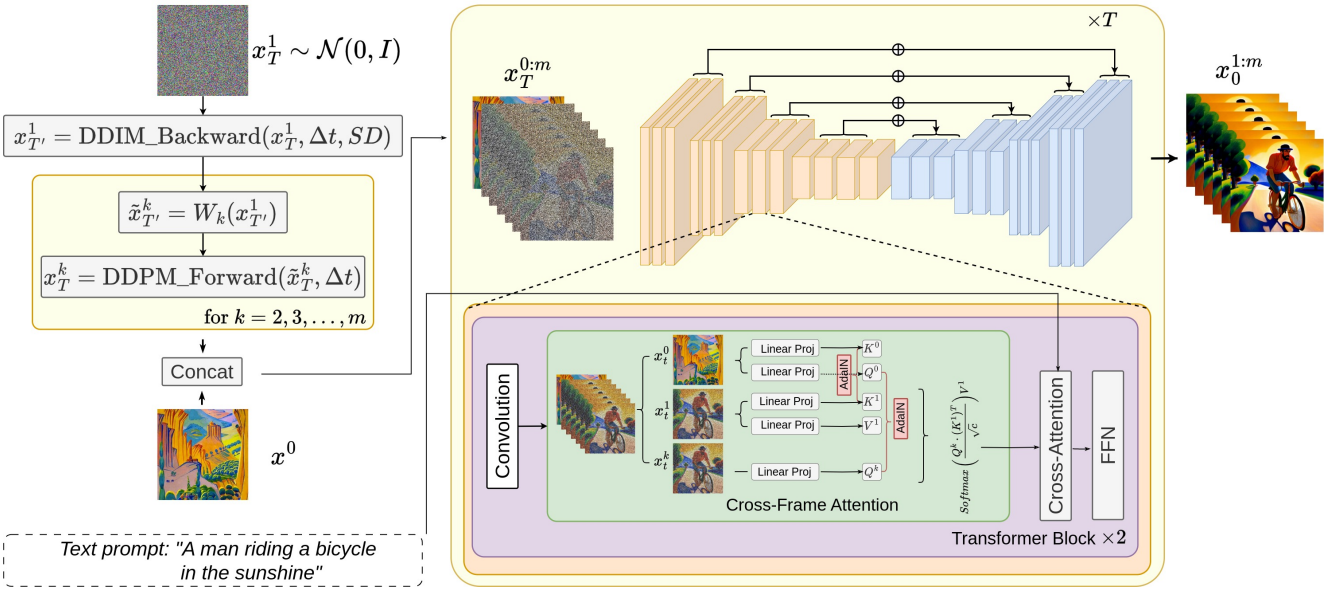

Figure 2. Architectural Overview of the StyleT2V0 Method: This diagram illustrates the integration of the shared attention mechanism within the Text2Video-Zero framework to enhance style consistency across video sequences. The method employs Adaptive Instance Normalization (AdaIN) to adjust the query $(\hat{Q}_i)$ for each frame F_i by blending style features from a designated reference image (Q_r), ensuring stylistic uniformity and alignment with textual prompts. Keys (K) and values (V) are also adapted using AdaIN, aligning them with the style of the reference image while maintaining content continuity from the first frame $(K1 \text{and} V1)$. The resultant attention mechanism applies these style-consistent queries, keys, and values to produce a video sequence where each frame is coherent in both style and content.

In addition to these mechanisms, Text2Video-Zero optionally applies background smoothing to enhance the visual quality of the video. Through salient object detection, a mask M_k is generated for each frame k to identify foreground pixels distinctly. A convex combination of the latent code x_1^t from the first frame warped to match the current frame k and the current frame's latent code x_k^t is used to refine the background.

$$x_k^t = M_k \odot x_k^t + (1 - M_k) \odot \left(\alpha \hat{x}_k^t + (1 - \alpha) x_k^t\right), \quad (6)$$

This method ensures that the background remains consistent and unobtrusive, allowing foreground elements to stand out and maintain narrative focus.

4. Method

In StyleT2V0, we extend the Text2Video-Zero framework by embedding a novel shared attention mechanism tailored to enhance style consistency across video sequences. This approach integrates style features from a designated reference image, ensuring uniform style across all frames while aligning with textual prompts. This is achieved through a specialized shared attention mechanism that utilizes Adaptive Instance Normalization (AdaIN) to blend style and content features across the sequence. The overview of our method can be found in Fig. 2.

The style image is consistently referenced to dictate the overarching style, while the content from the first frame serves as a baseline for content continuity.

For each frame F_i (where $i = 1, \ldots, n, $), the query Q_i is adjusted using AdaIN to incorporate style features from the style image:

$$\hat{Q}_i = \text{AdaIN}(Q_i, Q_r), \quad (7)$$

where Q_r is derived from the style image, effectively embedding its stylistic attributes into the current frame's query.

The keys K and values V are derived using AdaIN to align with the style properties of the style image, while maintaining content integrity from the first frame:

$$K = \text{AdaIN}(K_1, K_r), \quad V = V_1, \quad (8)$$

where K_1 and V_1 originate from the first frame, ensuring that the content features are consistently propagated through the video sequence. K_r is derived from the style image, aligning the keys with the desired style.

The attention for each frame F_i is computed to ensure that both style and content are coherently blended:

$$\text{Attention}(\hat{Q}_i, K, V) = \text{softmax}\left(\frac{\hat{Q}_i K^T}{\sqrt{d_k}}\right) V \quad (9)$$

where d_k is the dimensionality of the key vectors.

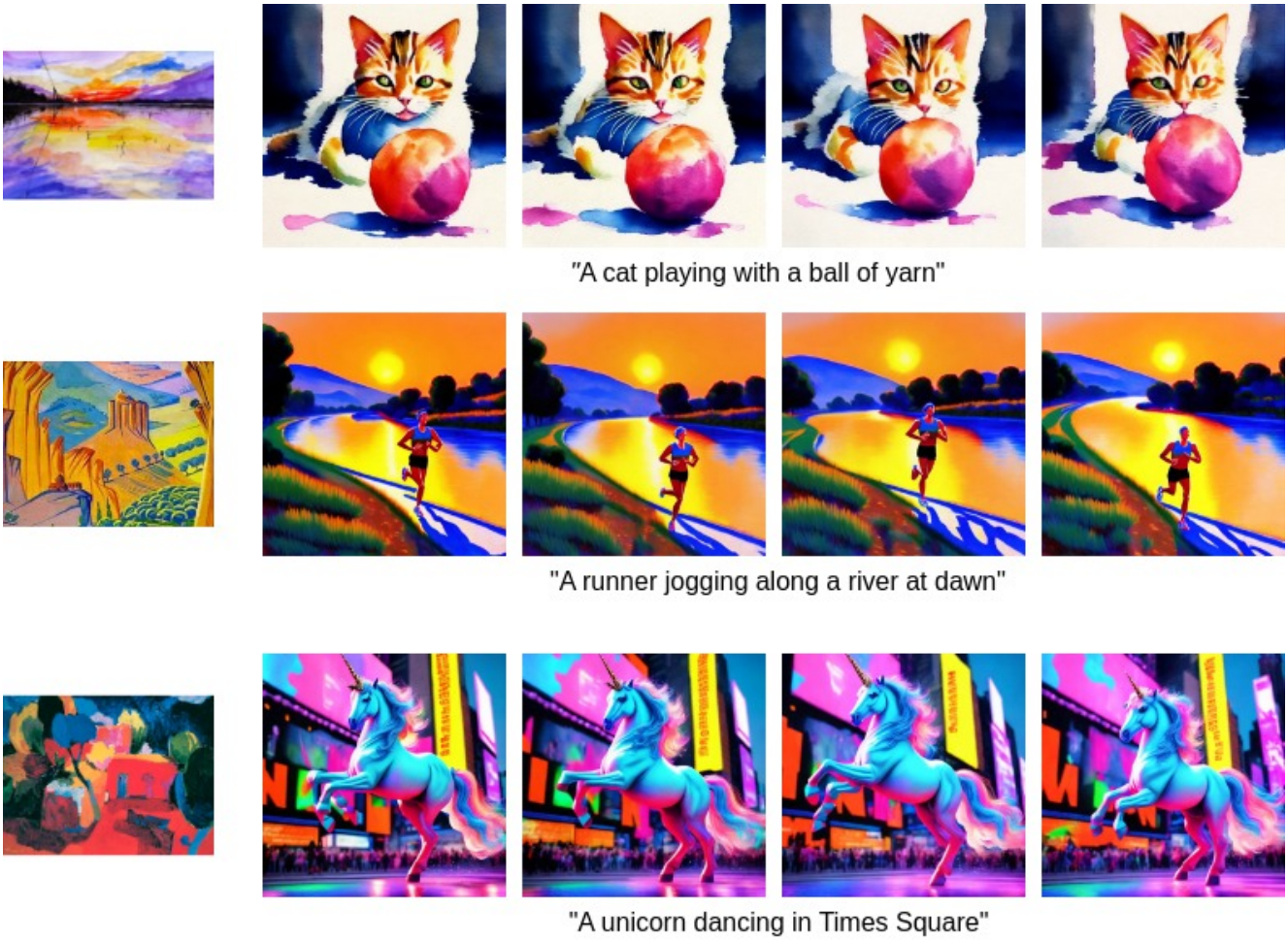

Figure 3. Results of StyleT2V0: The depicted frames demonstrate consistent styling and narrative fidelity, seamlessly integrating the style image and text prompt across the video sequence.

In addition to the shared attention mechanism, StyleT2V0 incorporates ControlNet [34] for detailed management of pose, edge, and depth attributes, enhancing the video's motion dynamics and visual clarity. Pose control ensures accurate character movements through joint information, while edge control sharpens scene definitions, and depth control introduces a layered spatial perception for immersive depth effects.

5. Experiments

5.1. Implementation Details

Our implementation is based on the publicly available Text2Video-Zero codebase [1], which we have adapted to integrate our methodological enhancements. These modifications include adjustments to the model's handling of latent space dynamics and attention mechanisms. Importantly, our approach extends the input modality to not only include text prompts but also reference images, allowing for a richer context in video generation.

[1] https://github.com/Picsart-AI-Research/Text2Video-Zero

5.2. Qualitative Results

The StyleT2V0 framework showcases a significant advancement in generating stylistically consistent short videos directly from textual prompts (see Fig. 3). The framework's flexibility was tested across a range of styles and contexts, showing its capability to adapt to various artistic and realistic styles. Whether translating a simple day-to-day activity or a complex, abstract concept into video format, StyleT2V0 maintained high fidelity to the original style and context of the reference image. Additionally, our method can incorporate edge, pose, and depth guidances (see Fig. ??) alongside the text and style image to enhance the visual coherence and dynamic representation of the videos.

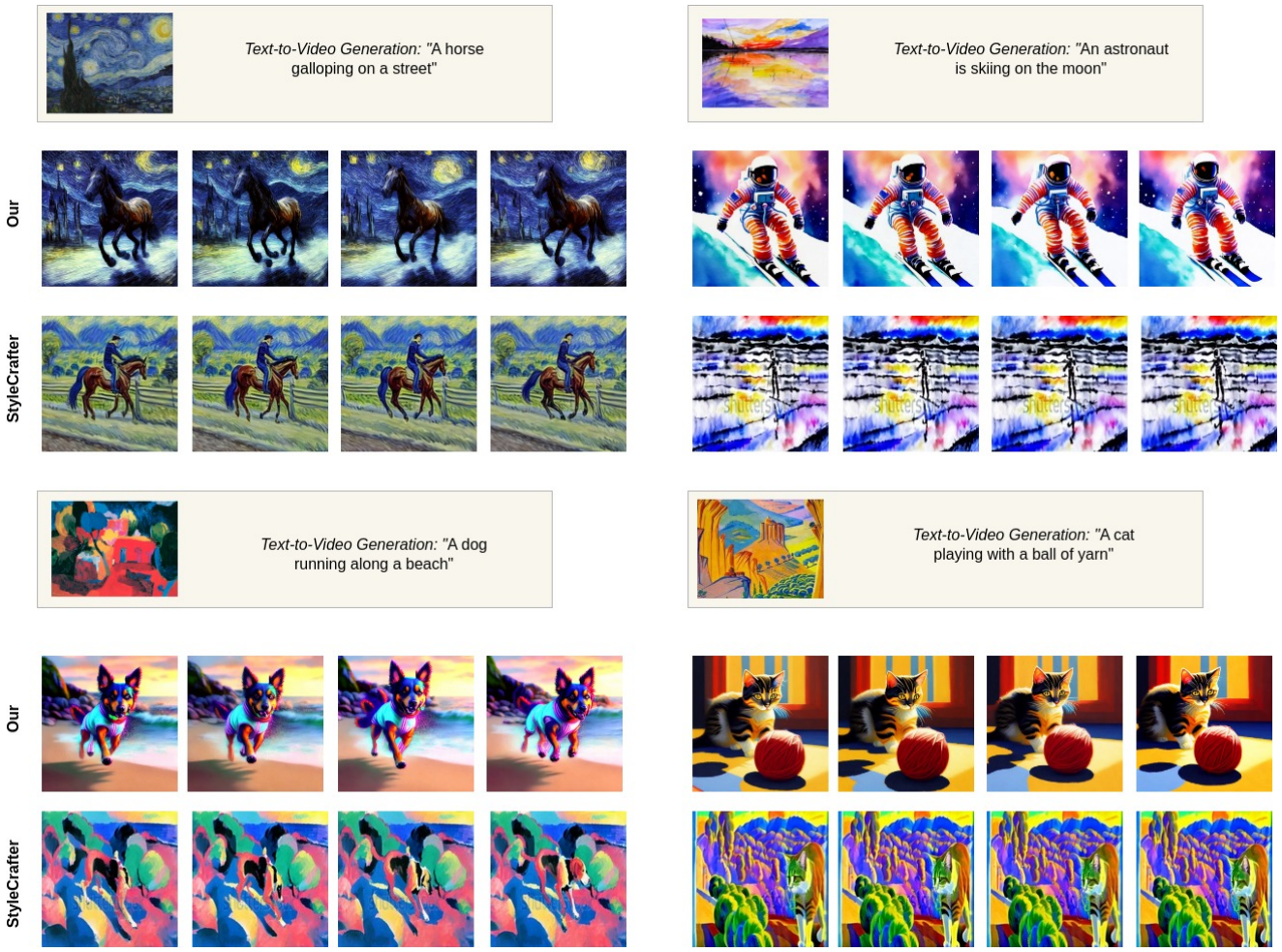

Figure 4. Comparision of StyleT2V0(our) with StyleCrafter.

In summary, StyleT2V0 not only advances the technical capabilities of text-to-video synthesis but also enhances the creative possibilities, allowing users to produce rich, stylistically consistent videos directly from textual descriptions. This marks a notable progress in the field of generative media, paving the way for more dynamic and aesthetically coherent video content generation.

5.3. Quantitative Results

Testing Dataset. To assess the effectiveness and scalability of our method, we assembled a test dataset comprising content prompts and style references. The content prompts consist of recognizable textual descriptions generated by GPT-4 [19], categorized into four meta-groups: human, animal, object, and landscape. For style references, we curated a collection of 10 diverse single-reference stylized images, encompassing both renowned artworks and generic style images. This structured dataset allows for a comprehensive evaluation of our video generation model across a variety of content and stylistic contexts.

Metrics. For the evaluation of our method, we focus on three key aspects: style alignment, text-to-video alignment, and video consistency. To ensure a comprehensive evaluation, we employ a combination of SSIM [30] and CLIP-based [10] metrics along with MAWE.

To assess style alignment and ensure consistent style within frames, we use the SSIM (Structural Similarity Index Measure) metric. SSIM evaluates the structural similarity between the style image and the video frames by comparing luminance, contrast, and structure. It quantifies how well the style's structure and texture are preserved across frames, ensuring that the video maintains the intended artistic style without significant deviations.

For text-to-video alignment, we employ the CLIP (Contrastive Language–Image Pre-Training) text score. CLIP is a model trained on a large dataset of images and their corresponding textual descriptions, allowing it to understand and compare visual and textual information effectively. The

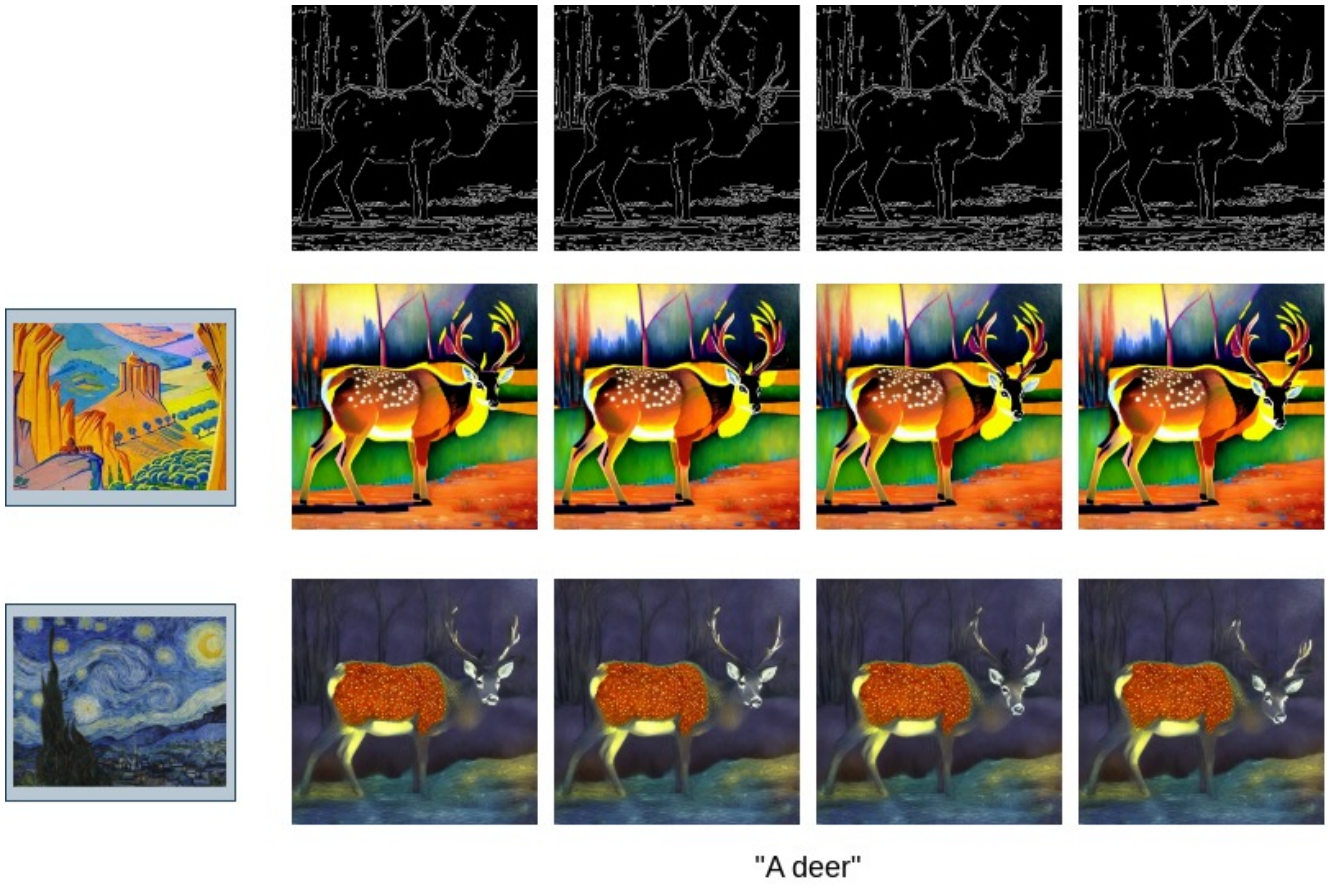

Figure 5. Conditional generation with edge control.

CLIP text score measures the similarity between the textual description provided as input and the visual content of the generated video. This metric ensures that the video accurately reflects the input text prompt, capturing the intended actions, objects, and settings described in the text.

Lastly, video consistency is evaluated using the Motion Aware Warp Error (MAWE), introduced in StreamingT2V [8]. MAWE measures both the amount of motion and the optical flow warp error in a generated video, yielding a low value when the video demonstrates substantial motion while maintaining consistency throughout. This metric ensures that objects, characters, and scenes maintain coherence throughout the video sequence, helping to identify any abrupt changes or inconsistencies in the visual content that might disrupt the viewer's experience.

Together, these metrics—SSIM, CLIP text score, and MAWE—provide a comprehensive evaluation framework for our method, ensuring that the generated videos are stylistically consistent, accurately aligned with the input text, and temporally coherent across frames.

	SSIM ↑	CLIP Text Score ↑	MAWE ↓
StyleCrafter	0.112	0.188	29.043
Only Prompt	0.121	0.174	6.277
StyleT2V0 (*Ours*)	0.123	0.195	7.958

Table 1. Quanitative comparison with Ablation Study and StyleCrafter methods. Best performing metrics are highlighted in red.

5.4. Comparison with Existing Models

We compare our method with StyleCrafter[16], a publicly available text-to-video model enhanced for style fidelity using a style control adapter. StyleCrafter is evaluated for its ability to generate stylized videos that align closely with user-provided reference images, making it a relevant benchmark for our approach.

5.4.1 Quantitative Comparison

We provide a detailed quantitative comparison Table 1 between our method and the StyleCrafter method using three evaluation metrics: SSIM (Structural Similarity Index Mea-

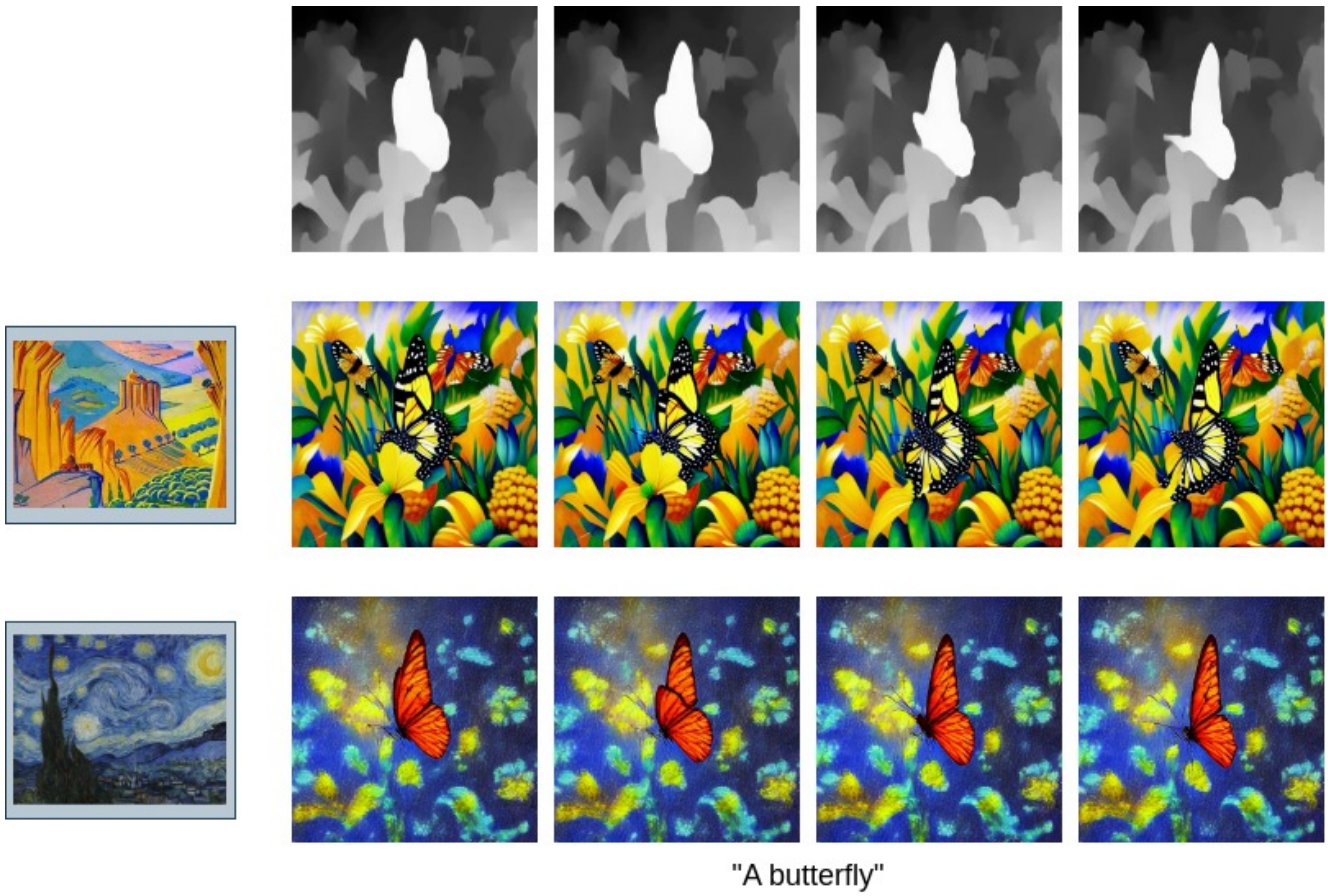

Figure 6. Conditional generation with depth control.

	Style Alignment	Text Alignment	Overall Quality
StyleCrafter	57.14%	22.73%	28.88%
StyleT2V0 (*Ours*)	42.86%	77.27%	71.12%

Table 2. Human Evaluation results for model comparison.

sure), CLIP Text Score, and MAWE (Mean Absolute Warping Error). Our results show that StyleT2V0 consistently outperforms StyleCrafter across all metrics. StyleT2V0 demonstrates superior structural integrity, better alignment with textual prompts, and significantly improved temporal consistency with less warping. Overall, StyleT2V0 proves to be a more effective method for style-aligned video generation.

We also conducted a human evaluation for the model comparison between our method and StyleCrafter. The evaluation focused on three main criteria: Style Alignment, Text Alignment, and Overall Quality. The specific questions asked to the evaluators were:

1. Which option has better style alignment?
2. Which option has better text alignment?
3. Which option is better overall?

The quantitative results of this evaluation are detailed in Table 2

The results indicate that while StyleCrafter performs better in style alignment, our method significantly outperforms it in text alignment and overall quality. This suggests that our approach is more effective in maintaining the alignment with the text and providing a better overall outcome, which is crucial for practical applications where both style and content accuracy are essential.

5.4.2 Qualitative Comparison

We present several results of our method in Fig. 8. and provide a qualitative comparison to StyleCrafter [16]. As can be seen from the results, there are cases when the style-video alignment is better preserved in StyleCrafter, but text-video alignment is better in our method, and vice versa. However, in general, it can be noted that our method better preserves both style-video and text-video alignment.

For example, in the video with the prompt ”A horse gal-

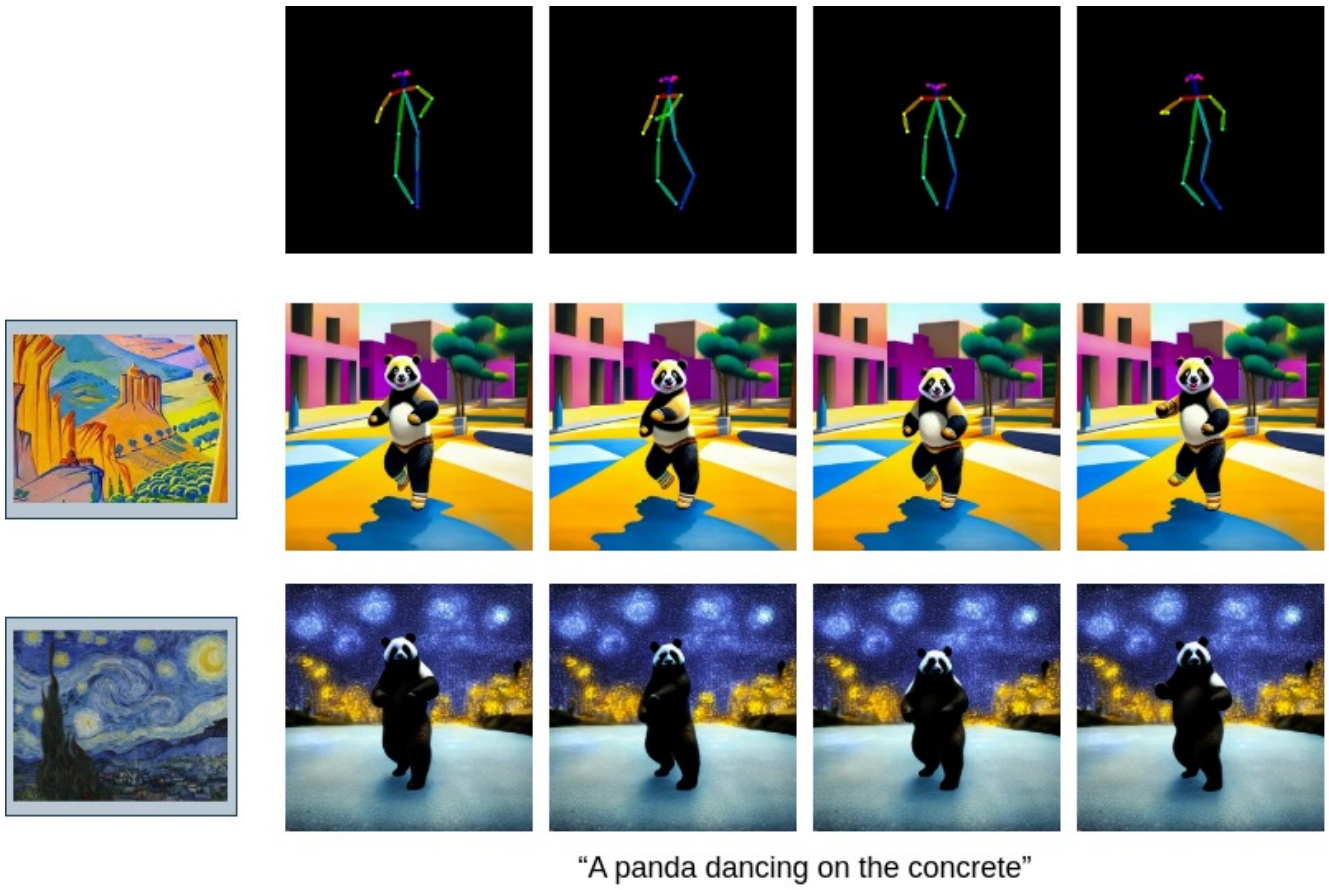

Figure 7. Conditional generation with pose control.

loping on a street" and the style image "Van Gogh's Starry Night," the style features are better preserved in our method. In the example with the prompt "A dog running along the beach" with the style image "Minas Avetisyan's Jajur," even though the colors and features from the style image seem to be better aligned in StyleCrafter, the objects and motions described in the textual prompt are not noticeable in the generated video, making it difficult to discern the dog in the video frames. This pattern is similarly observed in the other two examples provided.

5.5. Ablation Study

In our ablation study, we evaluated the model's performance by embedding style descriptions directly within the text prompts and by using both the style description and the style image, as opposed to leveraging an external style image alone. The first setup was designed to test the model's capability to infer and render stylistic elements based solely on textual descriptions, particularly assessing its effectiveness with recognizable and easily describable styles. We also experimented with using both the style image and style guidance in the text prompt. However, this approach did not yield notable results beyond those achieved with the style image alone. Our findings indicate that while the model can interpret and generate videos aligned with well-known styles described in the text, it faces considerable challenges when tasked with personal or obscure styles (see Figure 9). These less recognizable or describable styles are difficult for the model to accurately capture and reproduce without visual cues. This underscores the critical role of direct style inputs in enhancing the model's performance, particularly in applications where unique or customized stylistic fidelity is paramount. Overall, the ablation study highlights the importance of style representation in the model's input, demonstrating that while text-based style descriptions can suffice for general styles, they are insufficient for capturing the full spectrum of stylistic nuances required for high-quality, personalized video generation. Quantitative results can be found in Table 1.

6. Conclusion

In this study, we introduced StyleT2V0, a zero-shot framework for generating stylistically consistent videos from tex-

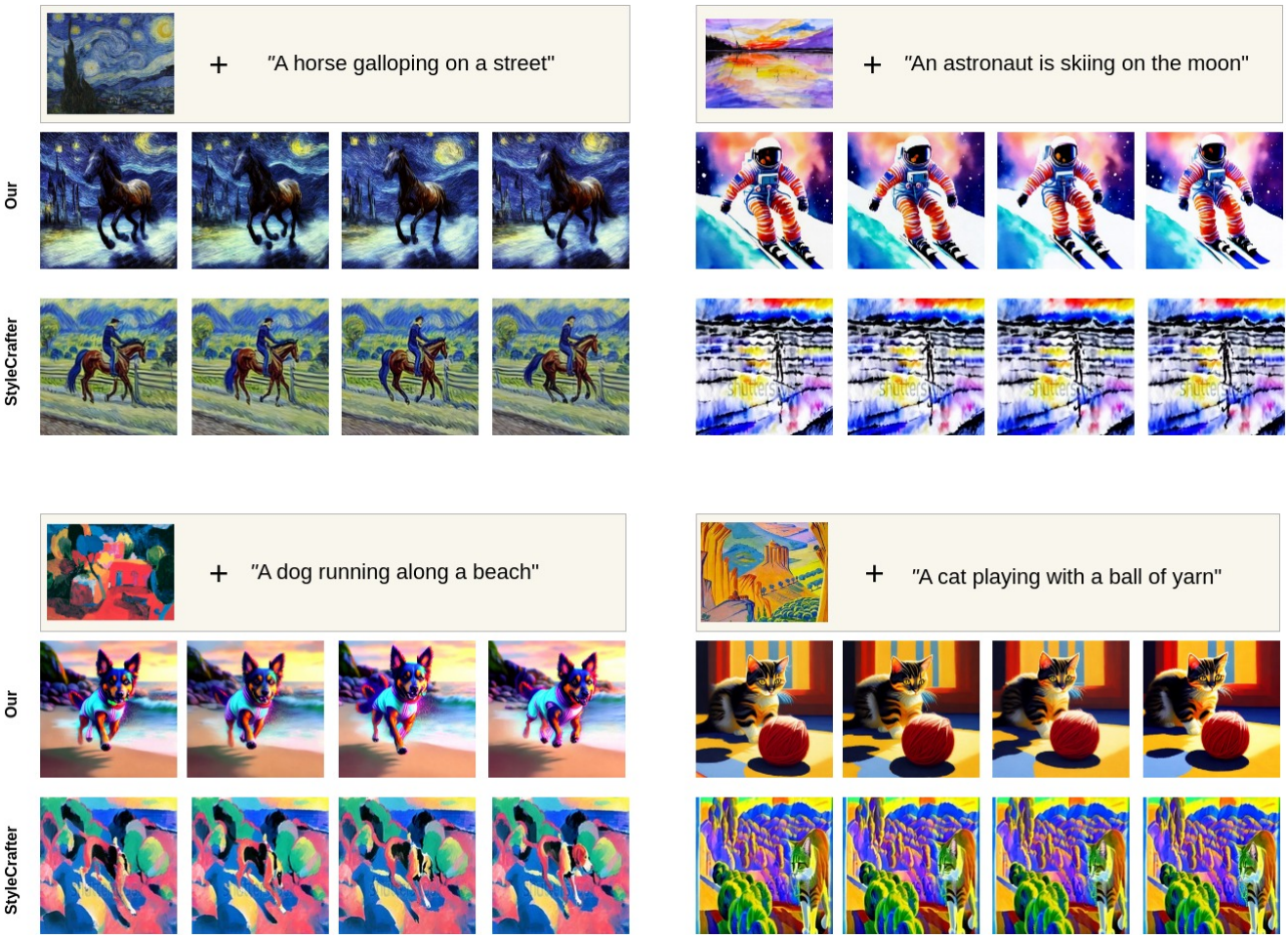

Figure 8. Comparision of StyleT2V0(our) with StyleCrafter.

tual prompts, aligned with specified reference images without the need for retraining. Our approach leverages shared attention mechanisms, ensuring that each video frame not only adheres to the contextual details of the text but also harmonizes with the aesthetic qualities of the reference image. Furthermore, the integration of pose, edge, and depth controls allows for precise manipulation of dynamic movement, visual sharpness, and spatial depth, enhancing the expressiveness and realism of the generated videos.

The experimental results validate StyleT2V0's ability to produce videos that are temporally consistent and stylistically faithful to both the textual prompts and the visual style of the reference images. Additionally, the adaptability to various control settings demonstrates the framework's versatility and its potential application across various domains such as digital marketing and personalized media production, significantly reducing the need for computational resources typically associated with training advanced video generation models.

References

[1] Openai's sora. https://openai.com/sora/. Published: 2024-02-15. 3

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 4

[3] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers, 2022. 3

[4] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models, 2023. 3

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2, 3

[7] Mark Hamazaspyan and Shant Navasardyan. Diffusion-

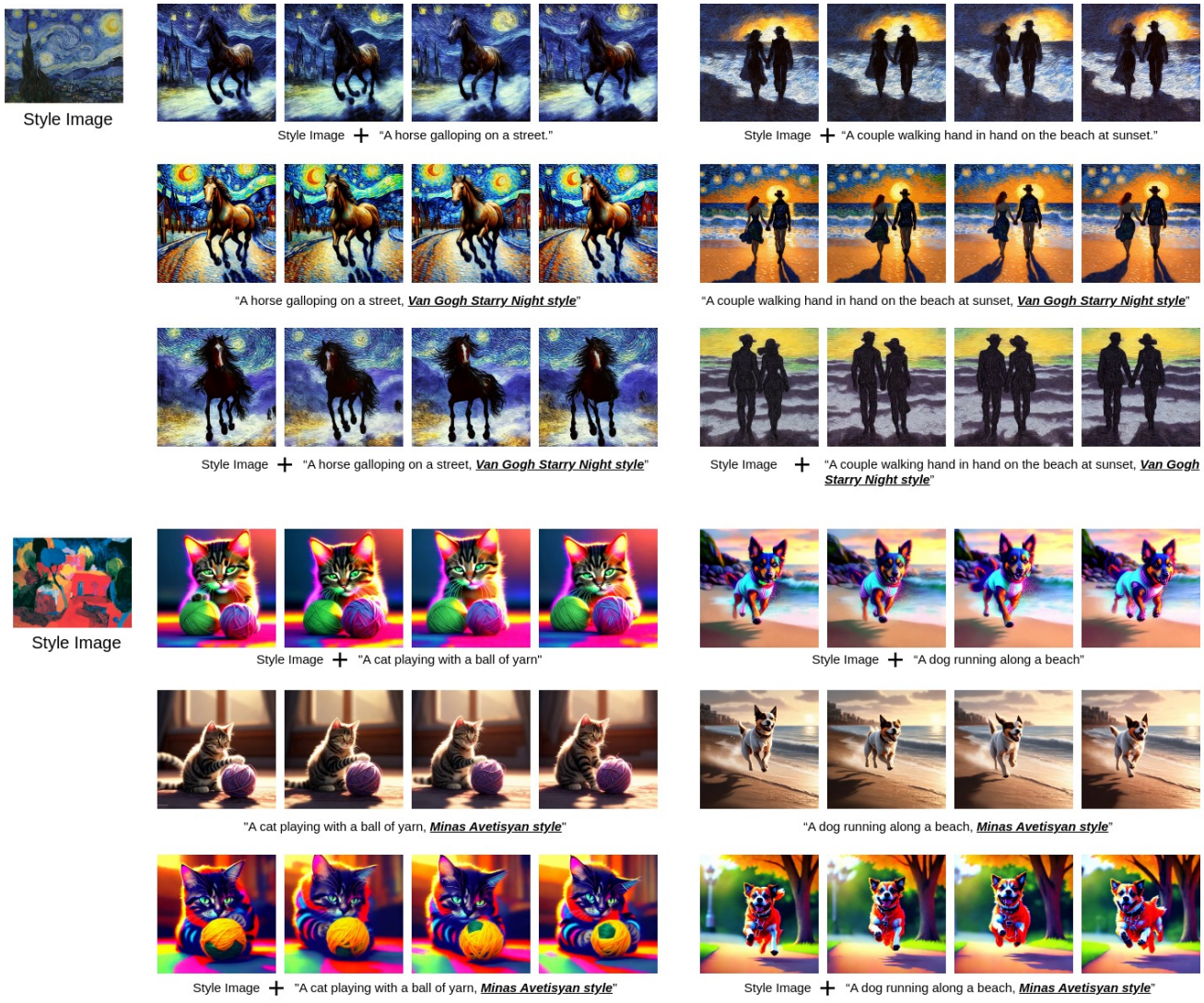

Figure 9. Ablation study showing the effect of our proposed style guided text-to-video method.

enhanced patchmatch: A framework for arbitrary style transfer with diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 797–805, 2023. 3

[8] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 8

[9] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention, 2024. 3

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 7

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 4

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022. 3

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 3

[14] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers, 2022. 3

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-

to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 3, 4

[16] Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Xintao Wang, Yujiu Yang, and Ying Shan. Stylecrafter: Enhancing stylized text-to-video generation with style adapter, 2023. 2, 3, 8, 9

[17] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2024. 3

[18] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022. 3

[19] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 7

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 4

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 3

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[24] 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. 2022. 3

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[26] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An,

Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 3

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 4

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3, 4

[29] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description, 2022. 3

[30] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 7

[31] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation, 2021. 3

[32] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 2

[33] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks, 2017. 2

[34] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 6