Text To Video Enhancing Video Generation Using Diffusion Models And Reconstruction Network

by Jiayao Jin

Submission date: 03-Sep-2023 01:02AM (UTC-0500)
Submission ID: 2156605012
File name: datacenter_paper_turnitin_2023-09-03_16222338.docx (1.44M)
Word count: 3820
Character count: 22690

Text To Video: Enhancing Video Generation Using Diffusion Models And Reconstruction Network

1st Jiayao Jin
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
1193210320@stu.jiangnan.edu.cn

2nd Jianhang Wu
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
1193210318@stu.jiangnan.edu.cn

3rd Zhoucheng Xu
School of Artificial Intelligence and Computer Science
Jiangnan University
Wuxi, China
1193210418@stu.jiangnan.edu.cn

4th Hangzhang
School of Machanical Engineering
Jiangnan University
Wuxi, China
1046210112@stu.jiangnan.edu.cn

5th Yaxin Wang
School of Textile Science and Engineering
Jiangnan University
Wuxi, China
1091210416@stu.jiangnan.edu.cn

6th Jielong Yang*
School of Internet of Things Engineering
Jiangnan University
Wuxi, China
jyang@jiangnan.edu.cn

***Abstract*—This paper proposes a method to improve the quality of generated videos in text to video generation techniques based on diffusion models, which suffer from low quality and poor continuity.The method involves dynamically adjusting the noise frame connections to enhance the video quality. A Reconstruction Net is introduced to automatically adjust the noise correlation among frames during the training process. Experimental results demonstrate that this method can enhance the quality of generated videos, improve video continuity, enhance the representation of image details, and strengthen the correspondence between generated and original videos. This research is of significant importance in advancing the development of text-based video generation techniques based on diffusion models.**

Keywords-Text to video; Diffusion model; Noise frame connections; Reconstruction Net;

I. INTRODUCTION

In this study, our objective is to generate a video that is consistent in both temporal and spatial aspects based on text input. Additionally, we aim to enhance the diversity of the generated videos to cater to user preferences.

However, text-to-video generation faces several challenges. Firstly, training a text-to-video model with high generalization requires a large amount of video training data, which can be time-consuming and resource-intensive. Secondly, the generated video frames often lack consistency, resulting in less smooth video transitions. Additionally, the quality of the generated videos falls short of professional production standards. Lastly, existing text-to-video models have limited reliance on the input text, struggling to accurately capture the semantic information contained in the text descriptions.

Several existing works have made progress in the field of text-to-video generation. For example, CogVideo[1] is capable of generating videos that are semantically aligned with the input text, but suffers from poor video quality. Plug-and-Play enables independent editing of each frame, but lacks consistency between frames. Text2LIVE[2] generates smooth and consistent videos, but exhibits limited reliance on the input text. TuneAVedio[3] partially addresses these issues, but still falls short in terms of video consistency.

To address the aforementioned challenges, we propose a novel model consisting of the following key components:

- Reconstruction Network(RCN): We introduce a latent space frame reconstruction network that learns the correlations between frames in the video's latent space. This network generates an alpha matrix, which is used to weight and enhance the associations between frames in the latent space of the video.
- Text Expansion Mechanism: We propose a text expansion mapping mechanism to ensure that the text embeddings of each video frame are not identical but related. This mechanism helps to increase the diversity of the generated videos by introducing variations in the textual descriptions associated with each frame.

In summary, our research focuses on enhancing the quality and continuity of text-to-video generation, while also expanding the diversity of the generated videos. The model has demonstrated significant improvements in the quality and continuity of the generated videos, which will contribute to advancing research and applications in this field.

II. RELATED WORKS AND BACKGROUND

Our work involves the following fields: diffusion models for generating images/videos from text prompts, text-based editing of an actual image or video, and generative models trained on a single video. Here, we briefly outline the main achievements in each field,

highlighting the similarities and distinctions between their approach and our proposed method.

A. *Text-to-Image diffusion models*

Many T2I models have been proposed[4][5][6][7]. DALL-E 2[8] proposes a two-stage model: the model leverages CLIP to generate image embeddings from text captions. The decoder then generates images conditioned on these embeddings, resulting in diverse samples while maintaining high levels of photorealism and caption similarity with minimal loss. Diffusion models are employed for computational efficiency and high-quality image synthesis. LAFITE[9], proposes a method trains T2I models without text data by utilizing the CLIP model's multimodal semantic space. It achieves cutting-edge performance in text-to-image tasks, outperforming models trained on complete text-image pairs. The language-free model can be fine-tuned, saving training time and cost while maintaining competitive performance with a significantly smaller size compared to the DALL-E model. Stable Diffusion[10], highlights the use of diffusion models (DMs) for image synthesis, offering state-of-the-art results and guiding mechanisms. However, optimizing these models in pixel space is computationally intensive, so the authors propose training them in the latent space of pretrained autoencoders to conserve resources while maintaining quality and flexibility. By introducing cross-attention layers, the resulting latent diffusion models (LDMs) excel in tasks like image inpainting, class-conditional synthesis, text-to-image synthesis, and super-resolution, surpassing existing methods while reducing computational requirements compared to pixel-based DMs.

B. *Text-to-Video generative models*

Many T2V models have also been proposed[11][12], .Make-A-Video[13], proposes a method for text-to-video generation by leveraging paired text-image data and unsupervised video footage. It achieves faster training, does not require paired text-video data, and produces high-resolution and faithful videos using spatial-temporal modules. The approach sets a new state-of-the-art in text-to-video generation based on qualitative and quantitative evaluations. CogVideo[14] is a 9B-parameter transformer trained based on the pretrained text-to-image model, CogView2. A multi-frame-rate hierarchical training strategy is proposed to improve the alignment between text and video clips. Tune-A-Video[3], proposes One-Shot Video Tuning, a new T2V generation setting using a single text-video pair. Building on T2I diffusion models, it leverages the ability to generate still images representing verbs and extends it to achieve content consistency in generating multiple images. The approach incorporates Tune-A-Video with spatio-temporal attention and efficient one-shot tuning. However, the frame-to-frame continuity of the video generated by this model is poor and not smooth enough. Our work is mainly based on improving this model to generate smoother videos with high quality and continuity.

C. *Text-driven video editing*

Blended Diffusion[14], introduces a method for performing local edits on natural images using a combination of natural language descriptions and an ROI mask. It uses a pretrained language-image model and a denoising diffusion probabilistic model to generate natural-looking results and blend the edited region with the unchanged parts of the image. Text2LIVE[2], describes a method for zero-shot, text-driven appearance manipulation in natural images and videos. The objective is to edit the appearance of objects or augment scenes based on a target text prompt. Instead of directly generating the edited output, the method generates an edit layer (color+opacity) that is composited over the original input, allowing for constraint-based generation and fidelity to the original image.

D. *Generation from a single video*

VGPNN[15], questions the necessity of GANs for single-video generation and manipulation tasks, proposing a non-parametric baseline using classical space-time patches-nearest-neighbors approaches. This simple approach outperforms GANs in visual quality and realism, with significantly reduced runtime (from days to seconds). It can be easily scaled to Full-HD videos, setting a new benchmark for single-video generation and manipulation tasks.

E. *enoising diffusion probablilisti models (DDPMs[16])*

This model is inspired by nonequilibrium statistical physics. In physics diffusion is an ordered to disordered process, which is similar to adding noise to a picture constantly.The structure within the data distribution is systematically and flexibly destroyed by an iteratively propagating forward diffusion process, and then a backward diffusion process is learned,which will restore the structure in the data, producing a highly flexible and easy-to-handle data generation model.Here is the specific derivation process. The diffusion modeldiffers from other types of latent variable models in that the approximate posterior is a fixed Markov chain $q(x_t \mid x_{t-1})$ and thus can be written as

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}) \tag{1}$$

The data that increases Gaussian noise $p(x_T) = N(x_T; 0, I)$ is controlled by beita is a variance schedule,which ranges from zero to one,and β_1 .Since the forward process is a fixed chain and therefore β_T also fixed, the actual code can be obtained through a fixed table. From global considerations, the noise added to the closer the data distribution, the more it affects the data distribution, and vice versa. Thus, it can be defined as:

$$q(x_{1:T} \mid x_0) = N(x_t : \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathrm{I}) \tag{2}$$

$$q(x_t \mid x_0) = N(x_t : \sqrt{\bar{a}}\, x_0, (1-\bar{a})\mathrm{I}) \tag{3}$$

We need to learn a model p_θ (Likelihood function) to get the probability distribution of the whole dataset.

$$p_\theta(x_0:T) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t) \quad (4)$$

$$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \sum_\theta(x_t,t)) \quad (5)$$

At last we will get

$$\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\varepsilon_\theta(x_t,t)) \quad (6)$$

As a result, these models can be interpreted as a sequence of weight sharing denoising autoencoders $\grave{o}_\theta(x_t,t)$, where $\grave{o}$ is an auxiliary independent random variable.The target can be simplified as

$$E_{x,\varepsilon\sim N(0,1),t}[\| \varepsilon - \varepsilon_\theta(x_t,t)\|_2^2] \quad (7)$$

F. Latent diffusion models(LDMs[10])

In the LDM model, a self-encoder ε and decoder are introduced. The training process involves two main phases. First, we train a self-encoder to create an efficient low-dimensional representation $(z = \varepsilon(x))$ space that closely approximates the data space. This avoids the need for excessive spatial compression, as we utilize diffusion models within the learned latent space, which scales better in terms of spatial dimensions. Second ,we train DDPM to eliminate noise from the sampled data. $c = \psi(P^*)$ is the embedding of textyal condition P^* .The objective is give by :

$$E_{z,\varepsilon\sim N(0,1),t,c}[\| \varepsilon - \varepsilon_\theta(z_t,t,c)\|_2^2] \quad (8)$$

III. METHODS

Figure 1 presents the general framework of our model. A video is processed through a Variational AutoEncoder[17] (VAE), which provides a latent space for each frame in the video. To reconstruct the continuity between each frame in the latent space, the latent spaces are processed through a RCNet to achieve a correlated latent space. Subsequently, a noise sequence, consistent with the shape of the video, is added, and the Unet predicts the noise.

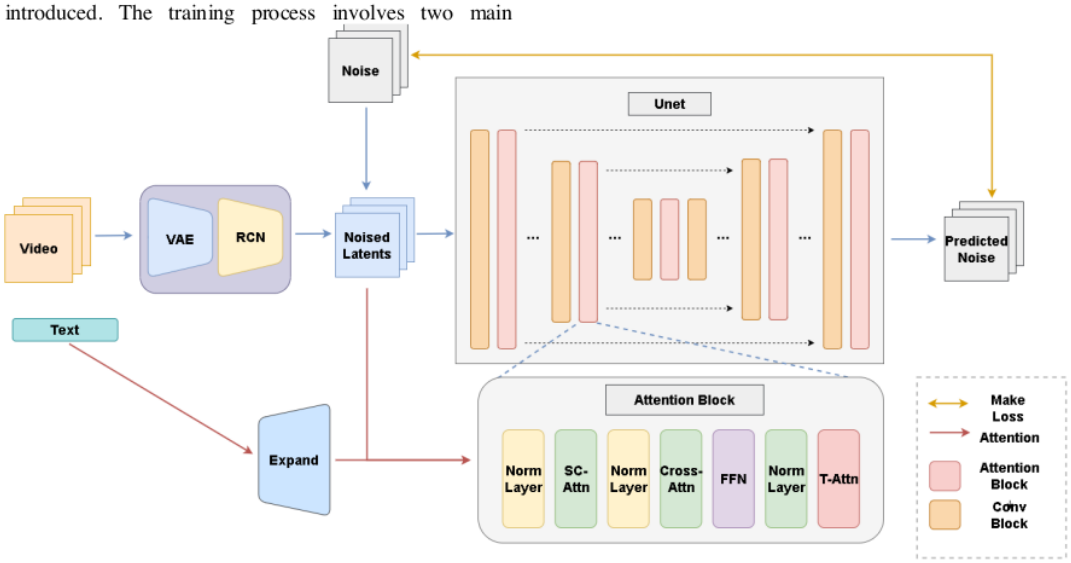


Figure 1. Main frames of VedioFromVedio

The attention module within the Unet, as shown in the figure, includes spatio-temporal attention[18], cross-attention, and self-attention[19]. The spatio-temporal attention focuses on the relationship between different frames in both space and time. The cross-attention enables the model to focus on crucial parts of the input when generating the output. This module enhances the model's ability to generate videos that accurately reflect the input text.

A. Reconstruction Network

To inflate text-to-image generation models (stable diffusion models) into text-to-video generation models while preserving the original model's high quality, the VAE in the T2I model is directly utilized to reduce the dimension of the original image to generate latent space vectors. The VAE can effectively enable each frame in the video to generate a latent space vector. However, we observed that even if there is strong consistency between video frames in pixel space, the consistency between video frames in the latent space after passing through the VAE is weakened. This results in the generated video's consistency being less than ideal.

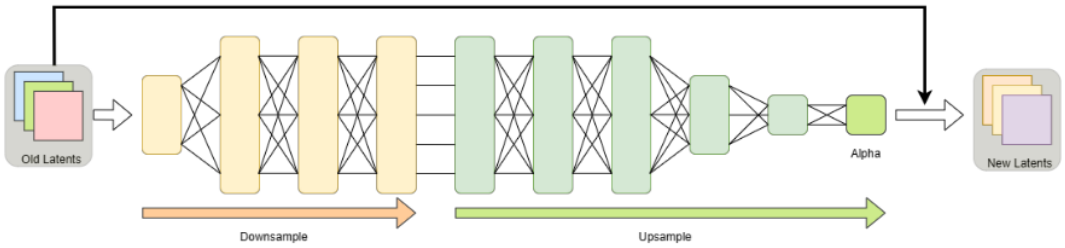


Figure 2. Reconstruction Network

Given the above problem, we propose a Reconstruction Network (RCN). This network can input the latent space of the video and output a video of the same shape after reconstruction. As shown in the figure 2, the network structure of the RCNet, within this network, the latent space learns the correlation matrix Alpha through multiple layers of 3D convolutional networks. Let's assume that the latents of a video are denoted by L , where L has a shape of $(1,c,f,h,w)$ .Here, c represents the number of channels in the latents, f represents the length of the video in frames, h represents the height of the latents, and w represents the width of the latents. We can express L as a set of individual frame latents: $L = \{L_i | 0 \le i \le f\}$, where L_i represents the latent representation of the i -th frame.

The formula for calculating the reconstructed latents (RL)is as follows:

$$RL_i = \begin{cases} \alpha \cdot L_{i-1} + (1-\alpha) \cdot X_i & \text{if } i > 0 \\ L_i & \text{if } i = 0 \end{cases} \quad (9)$$

Here, α is a relationship matrix trained by the RCNet (Reconstruction Network), which has a shape of $(1,b,1,h,w)$.

The network obtains a partial latent relationship, Alpha, between frames through downsampling and then upsampling. The new Latents are obtained by merging the original Latents with Alpha. The network continuously updates and trains to produce new Latents for subsequent training. The new Latents have a stronger connection between frames. This method enhances the consistency between video frames in the latent space, which will play a certain role in improving the performance of the model.

B. Word Embedding Alignment Strategy

In text-to-image generation models, a single word vector embedding is sufficient to encode an image. In Unet, noise prediction is assisted by embedding word vectors through CrossAttention. However, for text-to-video generation models, there is only one input text, but a whole video needs to be generated, meaning multiple word vectors are needed for auxiliary prediction. We observed that if the original word vectors are directly repeated to the current number of video frames, the video quality is not ideal. Therefore, we propose that when generating the word vector corresponding to each frame, a small amount of noise is added while ensuring the original word vector effect remains unchanged. This way, the word vector for each frame is consistent and differentiated.

C. Attention Mechanism

The attention mechanism is a method used to enhance the focus of neural network models on important parts of the input data. It is widely applied in fields such as natural language processing (NLP)[20][21][22] and computer vision (CV)[23][24].

In the attention mechanism, the input sequence (such as words in a sentence or pixels in an image) is represented as a set of vectors, where each vector is associated with weights. These weights indicate the contribution of each input element to the model's output.

By learning these weights, the attention mechanism can automatically select and focus on the most relevant and useful parts of the input sequence, thereby improving the model's performance.

The computation of attention weights, denoted as w , for given query vector q , key vector , and value vector v , can be achieved using the following formula:

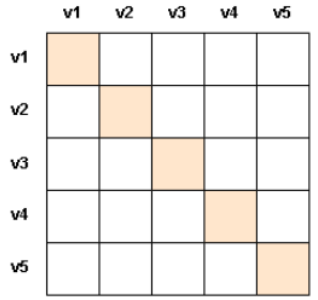
$$w = \text{softmax}\left(\frac{q \cdot k^T}{\sqrt{d_k}}\right) \quad (10)$$

Here, d_k represents the dimensionality of the query and key vectors. This formula first computes the dot product between the query vector q and the transposed key vector k , and then scales the result by dividing it by d_k . By applying the softmax function, the dot product result is transformed into attention weights w .

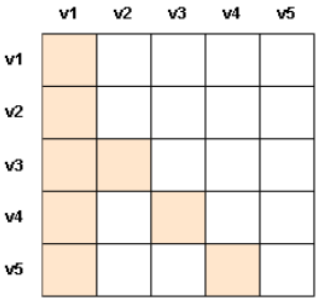
D. 3D Attention Mechanism

The Figure 3.a illustrates spatial self-attention. Consequently, the generated video appears to have good quality on a per-frame basis, but lacks temporal coherence between frames. Tune-A-Video[6], addresses the computational complexity of $O((mN)^2)$ associated with full attention and causal attention. Instead, it employs Sparse-Causal Attention(SC-Attn), as shown in the Figure 3.b, which incorporates the previous frame and the first frame as attention sources. This mechanism enables the generation of the current frame to be constrained by both the previous frame and the first frame, resulting in a higher consistency of the generated video.

Figure 4. Other result comparison.



(a) spatial self-attention



(b)spatio-temporal attention

Figure 3. Attention Block

In the T2I (Text-to-Image) model, attention mechanism is used for conditional embedding in multi-modal contexts such as generating text conditioned on images. In order to enhance the temporal consistency of the video, the attention module needs to be modified to the Attention Block module in Figure 1. Same as the TuneAVideo model, spatial self-attention is extended to sparse causal attention (SC-Attn). Additionally, CrossAttention is expanded to Cross3DAttention, which, similar to the T2I model, is used for conditional embedding. Self-attention is also extended to self-3D-Attention, and this attention mechanism is utilized for optimizing the network structure.

IV. Experiments

A. Dataset

We adopted the same approach as TuneAVideo[3], using a single commentary with text labels for training. Thanks to the high-quality pre-training model, our model only needs to learn the connection logic between the frames in this one video to generate a series of similar videos.

B. Experiment Setup

Parameter settings: We configured the 'channels_feature' in the reconstruction network as 10 and experimented with 'noise_scale' values of 1e-5, 1e-3, and 1e-2 in the text alignment strategy.

Baseline: While CogVideo[1] generates semantically related videos, their quality and conformity to conventional cognition are lacking. Plug-and-Play allows independent frame editing, but lacks consistency between frames. Text2LIVE[2] generates consistent videos, yet its text dependence is insufficient. TuneAVideo[3] improved some aspects, but video consistency and smoothness remain challenges. These methods serve as baselines, highlighting current research progress and challenges in text-to-video generation, providing insights and benchmarks for further development.

Evaluation: We comprehensively assessed our model-generated videos using subjective and objective methods for reliable results.

Subjective assessment: We evaluate the video based on its visual effect, consistency between content and text, fluency, creativity, content richness, and authenticity. Scores (0-100) were assigned using a questionnaire format. Equal-weighted indicators contributed 20% each to the final score.

Objective evaluation: To measure frame smoothness accurately, we computed average cosine similarity for each frame pair in the video.

C. Experimental Results

The results, as displayed in both Table 1 and Table 2, the results demonstrate that our model is more favored by users at the subjective level and, on the objective level, video consistency has improved. This indicates that our RCNet has effectively improved video quality and enhanced temporal consistency. Although there is a slight decrease in text-video alignment, it may be attributed to our model's strategy in handling word embeddings. Nevertheless, this also suggests that our strategy contributes to increased model diversity and innovation to a certain extent.

TABLE I. OBJECTIVE EVALUATION

Models	Frame Consistency	Text Consistency
CogVideo	90.64	23.91
Plug-and-Play	88.89	27.56
Tune-A-Video	92.40	**27.58**
Our Model	**93.52**	27.54

TABLE II. SUBJECTIVE ASSESSMENT

Model	Visual Quality	Content Consistency	Smoothness	Creativity & Diversity	Realism	Score
CogVideo	68.6	**90.5**	84.2	80.9	75.5	79.94
Plug-and-Play	82.2	81.3	88.0	85.7	80.4	83.52
TuneAVideo	88.4	89.3	90.5	87.9	90.2	89.26
Our Model	**92.3**	89.7	**92.9**	**92.7**	**91.7**	**91.86**

We conducted comparative experiments by controlling the noise_scale parameter under the same training settings to compare the models. First, we present the case where noise_scale is set to 1e-5. We generated videos using the trained prompt "(A wonder woman is running on the beach)". The two generated videos are unfolded frame by frame, as shown in Figure 5 below. As can be seen from the result, the clothing of Wonder Woman in our video is more consistent, the white clouds in the background are more realistic, and there are more reflections of Wonder Woman on the beach.

According to the comparison shown in the figure, the first row represents the results produced by the "Tune-A-Video" model, while the second row represents the results produced by our model. The overall effect is comparable, but our model demonstrates significant advantages in capturing details and constructing high-resolution output. Additionally, we tested our model with different types of prompts. The generated results for the prompts "(mickey mouse is skiing on the snow)" and "(Wonder Woman, donning a cowboy hat, engages in skiing.)" are shown in Figure 4. In the experiments, there are two sets of comparisons. The first row of each set represents the results from the original Tune-A-Video model, and the second row shows the results from our model.Observing the first set, it can be seen that the original model's handling of Mickey Mouse is not perfect. It even retains the yellow down jacket from the training video, which is greatly influenced by the training set. Our model performs better, both in terms of character detail and background detail. However, there's a shortcoming in the 5th frame where the character is missing an ear.For the second set in the Figure above, the hat in the original model is not stable. Although our model doesn't generate a hat, the overall quality of the deeper layers is relatively higher.

Figure 5. "A wonder woman is running on the beach" Result comparison

V. CONCLUSION AND DISCUSSION

We propose that the reconstruction network strengthens the correlation between video frames and adds a small amount of noise to each frame to enhance the consistency between text and video. Through experiments, it is shown that our method can improve Frame Consistency and Text Consistency.In the future, we will continue to improve the consistency of the video, and work on increasing the frame length of the generated video and improving the generalization of the model.

REFERENCES

[1] Hong, Wenyi, et al. "Cogvideo: Large-scale pretraining for text-to-video generation via transformers." arXiv preprint arXiv:2205.15868 (2022).

[2] Bar-Tal, Omer, et al. "Text2live: Text-driven layered image and video editing." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV. Cham: Springer Nature Switzerland, 2022.

[3] Wu, Jay Zhangjie, et al. "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation." arXiv preprint arXiv:2212.11565 (2022).

[4] S. Ye, H. Wang, M. Tan and F. Liu, "Recurrent Affine Transformation for Text-to-image Synthesis," in IEEE Transactions on Multimedia, doi: 10.1109/TMM.2023.3266607.

[5] Liao, Wentong, et al. "Text to image generation with semantic-spatial aware gan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] Huang, Yupan, et al. "Unifying multimodal transformer for bi-directional image and text generation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[7] Ramesh, Aditya, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning. PMLR, 2021.

[8] Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 (2022).

[9] Zhou, Yufan, et al. "Lafite: Towards language-free training for text-to-image generation." arXiv preprint arXiv:2111.13792 (2021).

[10] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[11] Balaji, Yogesh, et al. "Conditional GAN with Discriminative Filter Generation for Text-to-Video Synthesis." IJCAI. Vol. 1. No. 2019. 2019.

[12] Kim, Doyeon, Donggyu Joo, and Junmo Kim. "Tivgan: Text to image to video generation with step-by-step evolutionary generator." IEEE Access 8 (2020): 153113-153122.

[13] Singer, Uriel, et al. "Make-a-video: Text-to-video generation without text-video data." arXiv preprint arXiv:2209.14792 (2022).

[14] Avrahami, Omri, Dani Lischinski, and Ohad Fried. "Blended diffusion for text-driven editing of natural images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[15] Haim, Niv, et al. "Diverse generation from a single video made possible." Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII. Cham: Springer Nature Switzerland, 2022.

[16] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in neural information processing systems* 33 (2020): 6840-6851.

[17] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[18] Fu, Yang, et al. "Sta: Spatial-temporal attention for large-scale video-based person re-identification." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

[19] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[20] Galassi, Andrea, Marco Lippi, and Paolo Torroni. "Attention in natural language processing." IEEE transactions on neural networks and learning systems 32.10 (2020): 4291-4308.

[21] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

[22] Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation." arXiv preprint arXiv:1508.04025 (2015).

[23] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 7132-7141, doi: 10.1109/CVPR.2018.00745.

[24] Li, Xiang, et al. "Selective kernel networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

Text To Video Enhancing Video Generation Using Diffusion Models And Reconstruction Network

ORIGINALITY REPORT

21%	14%	16%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	arxiv.org Internet Source	2%
2	export.arxiv.org Internet Source	2%
3	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 Publication	2%
4	doiserbia.nb.rs Internet Source	1%
5	hosei.repo.nii.ac.jp Internet Source	1%
6	www.arxiv-vanity.com Internet Source	1%
7	es.slideshare.net Internet Source	1%
8	github.com Internet Source	1%

9	Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Bjorn Ommer. "High-Resolution Image Synthesis with Latent Diffusion Models", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 Publication	<1 %
10	Omri Avrahami, Dani Lischinski, Ohad Fried. "Blended Diffusion for Text-driven Editing of Natural Images", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 Publication	<1 %
11	Patrick P.K. Chan, Xiaotian Wang, Zhe Lin, Daniel S. Yeung. "Progressive editing with stacked Generative Adversarial Network for multiple facial attribute editing", Computer Vision and Image Understanding, 2021 Publication	<1 %
12	Ziwen Sun, Jiajie Liu, Zhicheng Ji. "Distributed Fusion Steganalysis Based on Combination System Likelihood Function", 2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 2011 Publication	<1 %
13	Submitted to Stony Brook University Student Paper	<1 %

14	Yan Luo, Zhichao Zuo, Zhao Zhang, Zhongqiu Zhao, Haijun Zhang, Richang Hong. "High-Fidelity Diffusion Editor for Zero-Shot Text-Guided Video Editing", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	<1 %
15	hdl.handle.net Internet Source	<1 %
16	Submitted to British University in Egypt Student Paper	<1 %
17	Submitted to BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA BIBLIOTECA Student Paper	<1 %
18	Submitted to Babes-Bolyai University Student Paper	<1 %
19	Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang et al. "Multimodal Image Synthesis and Editing: A Survey and Taxonomy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023 Publication	<1 %
20	Submitted to National University of Singapore Student Paper	<1 %
21	api.deepai.org Internet Source	<1 %

22	Chun Liu, Jingsong Hu, Hong Lin. "SWF-GAN: A Text-to-Image model based on sentence-word fusion perception", Computers & Graphics, 2023 Publication	<1 %
23	Submitted to Imperial College of Science, Technology and Medicine Student Paper	<1 %
24	Submitted to University of Sydney Student Paper	<1 %
25	Jinzhi Deng, Yan Wei, Jiangtao Liu. "Text-to-Image algorithm Based on Fusion Mechanism", Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence, 2022 Publication	<1 %
26	Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, Tali Dekel. "Chapter 41 Text2LIVE: Text-Driven Layered Image andVideo Editing", Springer Science and Business Media LLC, 2022 Publication	<1 %
27	Qunyan Jiang, Ting Rui, Juying Dai, Faming Shao, Guanlin Lu, Jinkang Wang. "A real-time detection method of multi-scale traffic signs	<1 %

based on dynamic pruning strategy", Multimedia Tools and Applications, 2023
Publication

28	amslaurea.unibo.it Internet Source	<1 %
29	cris.iucc.ac.il Internet Source	<1 %
30	ebin.pub Internet Source	<1 %
31	vdoc.pub Internet Source	<1 %
32	"Computer Vision – ECCV 2018", Springer Science and Business Media LLC, 2018 Publication	<1 %
33	Ali Köksal, Kenan E. Ak, Ying Sun, Deepu Rajan, Joo Hwee Lim. "Controllable Video Generation with Text-based Instructions", IEEE Transactions on Multimedia, 2023 Publication	<1 %
34	Chenyang Liu, Rui Zhao, Jianqi Chen, Zipeng Qi, Zhengxia Zou, Zhenwei Shi. "A Decoupling Paradigm with Prompt Learning for Remote Sensing Image Change Captioning", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	<1 %

35	Yun-Cheng Wang, Jintang Xue, Chengwei Wei, C.-C. Jay Kuo. "An Overview on Generative AI at Scale with Edge-Cloud Computing", Institute of Electrical and Electronics Engineers (IEEE), 2023 Publication	<1 %
36	Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao. "Vision-Language Pre-Training: Basics, Recent Advances, and Future Trends", Foundations and Trends® in Computer Graphics and Vision, 2022 Publication	<1 %
37	escholarship.org Internet Source	<1 %
38	lilianweng.github.io Internet Source	<1 %
39	proceedings.mlr.press Internet Source	<1 %
40	ro.uow.edu.au Internet Source	<1 %
41	scholar.archive.org Internet Source	<1 %
42	trepo.tuni.fi Internet Source	<1 %
43	www.hindawi.com Internet Source	<1 %

44	www.mdpi.com Internet Source	<1%
45	Jinkuan Zhu, Pengpeng Zeng, Lianli Gao, Gongfu Li, Dongliang Liao, Jingkuan Song. "Complementarity-aware Space Learning for Video-Text Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, 2023 Publication	<1%
46	Gerhard Paaß, Sven Giesselbach. "Foundation Models for Natural Language Processing", Springer Science and Business Media LLC, 2023 Publication	<1%
47	Shuo Yang, Xiaojun Bi, Jian Xiao, Jing Xia. "A Text-to-Image Generation Method Based on Multiattention Depth Residual Generation Adversarial Network", 2021 7th International Conference on Computer and Communications (ICCC), 2021 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		