TiV-ODE: A Neural ODE-based Approach for Controllable Video Generation From Text-Image Pairs

Yucheng Xu[1], Nanbo Li[1], Arushi Goel[1], Zonghai Yao[2], Zijian Guo[3]
Hamidreza Kasaei[4], Mohammadreza Kasaei[1], Zhibin Li[5]

***Abstract*— Videos capture the evolution of continuous dynamical systems over time in the form of discrete image sequences. Recently, video generation models have been widely used in robotic research. However, generating controllable videos from image-text pairs is an important yet underexplored research topic in both robotic and computer vision communities. This paper introduces an innovative and elegant framework named TiV-ODE, formulating this task as modeling the dynamical system in a continuous space. Specifically, our framework leverages the ability of Neural Ordinary Differential Equations (Neural ODEs) to model the complex dynamical system depicted by videos as a nonlinear ordinary differential equation. The resulting framework offers control over the generated videos' dynamics, content, and frame rate, a feature not provided by previous methods. Experiments demonstrate the ability of the proposed method to generate highly controllable and visually consistent videos and its capability of modeling dynamical systems. Overall, this work is a significant step towards developing advanced controllable video generation models that can handle complex and dynamic scenes.**

I. INTRODUCTION

Controllable video generation from image-text pairs aims to generate videos corresponding to given control signals, which allows for precise manipulation of various aspects of videos, such as appearance and motions. This level of controlled video generation is crucial for a wide range of applications, such as video editing, and custom video generation. Moreover, controllable video generation models have recently been used in edge-cutting robot research to synthesize data for model training [1], [2], [3], or to function as a robot planner [4], [5], [6]. However, compared to static images, videos have an additional temporal dimension to be modeled. The appearance and states of objects in the video are tightly coupled with the temporal dimension – the model must generate visually consistent content while predicting temporal changes based on motion cues to maintain motion consistency. Previous methods focus on generating controllable videos from images [7], [8], [9], [10] or from texts [11], [12]. However, Image-to-Video methods typically have no control over the motions in the generated videos, whereas Text-to-Video methods offer limited control over the appearance of generated videos. Hence, to facilitate control over both motion and appearance in video generation, it is necessary to combine image and text signals.

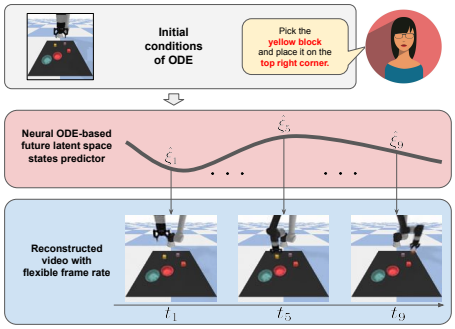

Fig. 1: An example of controllable video generation from a static image and a text caption using our proposed TiV-ODE. The underlying dynamical system is modeled using Neural ODE. Our model is capable of generating highly controllable and visually consistent video frames at any desired timesteps.

Another vital limitation of previous controllable video generation methods is the lack of modeling of the underlying continuous dynamical system from videos. The dynamical system refers to a closed system that governs changes of the entire environment (e.g. the dynamics of objects). In the predicted image sequences, the appearance and motion of observed objects shall always be consistent with given control signals – consistency between the underlying dynamical system and the generated videos. Prior methods typically model the underlying dynamical system as a discrete function of time, ignoring the fundamental difference between the continuous time dimension and the discrete image dimension as discussed in [13], [14]. Such approaches limit the ability to generate videos with flexible frame rates and handle videos with arbitrary frame rates. Moreover, in various applications, such as slow-motion video processing [15] or high-speed camera video processing [16], the regular timestep assumption does not hold. Therefore, a new controllable video generation method is needed which should be capable of generating highly controllable videos while correctly modeling the underlying continuous dynamical system.

To address these limitations, we developed a framework **T**ext-**i**mage-to-**V**ideo Ordinary Differential Equation (**TiV-ODE**). Firstly, our proposed method leverages the

[1] Yucheng Xu, Nanbo Li, Arushi Goel and Mohammadreza Kasaei are with the School of Informatics, University of Edinburgh, UK. Email: {yucheng.xu, nanbo.li, m.kasaei}@ed.ac.uk, goel.arushi@gmail.com

[2] Zonghai Yao is with the College of Information and Computing Sciences at the University of Massachusetts-Amherst, US. Email: zonghaiyao@umass.edu

[3] Zijian Guo is with the Department of Electrical and Computer Engineering, Boston University, US. Email: zjguo@bu.edu

[4]Hamidreza Kasaei is with the Department of Artificial Intelligence, Bernoulli Institute, University of Groningen, The Netherlands. Email: hamidreza.kasaei@rug.nl

[5] Zhibin Li is with the Department of Computer Science, University College London, UK. Email: alex.li@ucl.ac.uk

This work is supported by EU H2020 project Enhancing Healthcare with Assistive Robotic Mobile Manipulation (HARMONY, 101017008).

advantages of both Image-to-Video methods and Text-to-Video methods since images and texts are two complementary signals, static images provide rich visual information, while text captions describe the dynamic processes within videos in human language. By combining image input and text input, both the visual appearance and the physical motions within the videos can be further constrained to allow a higher level of control over the video content. Secondly, stemming from the physical modeling of dynamical systems [17], [18], Neural ODE [19] is incorporated in our proposed method to model the underlying continuous dynamical systems as ordinary differential equations (ODEs). By solving the ODE at arbitrary timestamps, our model is able to generate videos with flexible frame rates efficiently (See Figure 1). To the best of our knowledge, the proposed method is a new approach to solving controllable video generation problems. We summarize our contributions as follows:

- We proposed a novel video generation framework, TiV-ODE, which is capable of generating highly controllable and visually consistent videos conditioned on a single image and a text caption.
- Our proposed method is able to generate videos with flexible frame rates by leveraging Neural ODE to model the underlying continuous dynamical system from videos.
- We created a new dataset, the Synthetic Robot Pick-and-Place dataset – video sequences depicting a robot performing pick-and-place tasks with corresponding text captions – for evaluating our method and demonstrating its effectiveness. We also performed experiments on existing datasets such as CATER and Moving MNIST and showed improvements compared to previous works.

To the best of our knowledge, this is the first work that approaches the problem of controllable video generation from the dynamical system perspective, i.e., the system modeled as an ODE evolves according to *constraints set by a given initial condition*, resulting in a sequence of visual observations which form the generated video.

II. Related Works

A. Controllable Video Generation

A synthetic video can be generated in a number of ways using various conversion techniques. The controllable video generation methods that are most pertinent to our work include Image-to-Video and Text-to-Video methods.

Image-to-Video methods generate video sequences conditioned by given images. However, as a static image provides no motion clues, to facilitate video generation with editable scene dynamics, these methods require additional input to control the motions within the generated video, such as sparse trajectories [7], [14], and semantic masks [8], [9]. Existing Image-to-Video methods can only achieve low-level control of the generated videos, thus they are not suitable to be used to generate videos with complex motions [13].

Text-to-Video methods aim at generating video sequences from text captions. However, the appearance and motion information in the text caption is highly ambiguous leading to unavoidable uncertainties in generated videos. SyncDraw [11] is the first framework proposed to solve Text-to-Video tasks. Recently, GODIVA [12] was proposed to generate open-domain videos from given text captions. Given the ambiguous nature of the text, Text-to-Video methods can only achieve a low level of control over the generated videos. As a result, the appearance and motions within generated videos are mostly determined by the training dataset.

There is limited research work focused on combining the advantage of both Image-to-Video methods and Text-to-Video methods. To the best of our knowledge, the work in [20] is the closest one to our work. The work in [20] proposed a framework, MAGE, which generates videos from images with text captions. A motion embedding is used in MAGE [20] to memorize the motion patterns after observing the whole video, while our method formulates the underlying continuous dynamical system as an ordinary differential equation (ODE) and approximates it using a neural network [19]. Compared to MAGE, our method is able to generate controllable videos with flexible frame rates, which greatly widens its potential applications in robotics research. A detailed comparison between our method and the MAGE is presented in Section IV.

B. Dynamical System Understanding from Videos

Modeling and understanding dynamical systems from videos is important for video processing. Previous methods typically model the underlying dynamical system using an RNN-based structure [14], [21], [22], [23] or a transformer-based structure [24] that can represent the temporal information. Another energy-based Spatial-Temporal generative model was proposed in [25], [26], learning the dynamic patterns in video sequences by matching the synthesized signals from the sampled Langevin dynamics to the observed training signals. However, since these methods are mostly Video-to-Video methods, which are affected by the dynamics bias from the training data, they failed to generate videos with editable dynamics.

C. Neural ODE

Neural Ordinary Differential Equations (Neural ODEs, NODEs) [19] interprets the forward pass of a ResNet [27] as solving an ordinary differential equation. It is designed to model the temporal evolution of any dynamical system. Recent works [28], [29], [30] have shown the power of Neural ODE for modeling time series. Augmented Neural ODEs (ANODEs) [19] was proposed to extend the original Neural ODE by augmenting the latent space, which makes it a universal approximator [31], [32]. The method in [33] introduces the Neural ODE into video generation tasks to model time-continuous dynamics within the videos over a continuous latent space. Vid-ODE [13] further combines Neural ODE with GAN [34] to improve the quality of generated videos, however, Vid-ODE [13] relies on the input video clip to estimate the latent dynamics, which makes the generated videos hugely biased by the input video clips.

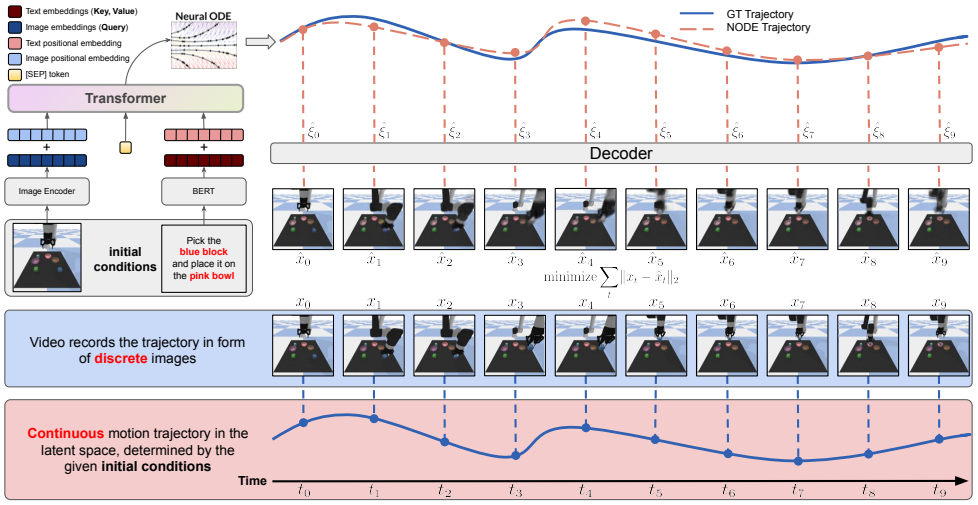

Fig. 2: System structure of TiV-ODE. The continuous physical motions of objects are recorded as a set of discrete video frames. The latent vectors of the video are assumed to follow an ODE trajectory that corresponds to the motions of objects (See the blue and red blocks). Given the initial image and the text caption, the whole video sequence is generated as follows: (1) the input image and text are encoded by the image encoder and text encoder respectively. (2) Together with positional embeddings, the image embeddings and text embeddings are fused by a transformer to generate a Text-image embedding. (3) The Text-image embedding is used as the initial condition of Neural ODE, then the Neural ODE is solved at the desired timesteps using a numerical ODE solver to generate latent vectors at every timestep. (4) The generated latent vectors are quantized by the codebook and decoded by the image decoder to generate video frames at every timestep. The training objective of our model is to minimize the distance between each pair of data points at each desired timestep.

III. TIV-ODE for Controllable Video Generation

In this section, we first explain how we formulate the problem of controllable video generation by learning the dynamical system using Neural ODE, followed by a discussion on the general architecture of our proposed TiV-ODE. Then, details of our TiV-ODE, including the VQ-VAE for image generation, the text-image fusion module, and the Neural ODE module, will be presented individually.

A. Problem Formulation

This paper targets the Text-image-to-Video task with modeling of the underlying continuous dynamical system. Let $x_t \in \mathcal{X} \subset \mathbb{R}^N$ be an image observation of the system (defined in the image sample space $\mathcal{X}$) at time point t, and $s \in S \subset \mathbb{R}^L$ be the text caption (defined in the text sample space S). We aim to model the dynamical system defined over the text-image domain ($\mathcal{X} \times S$) such that, given a text caption s and an image observation x_0 as the initial conditions, our model can generate the image observations x_t for any $t \geq 0$. Unlike the previous method [20] which models a dynamical process using a discrete state-transition, i.e. $x_t = \text{RNN}(x_{t-1}, s)$, we model the system as a continuous vector field, $(x_0, s) \mapsto x(t),\ \forall\ t \in (0, 1)$, that is saying we want to approximate a function $\dot{x}(t) = F(x_0, s, t), \forall t \in (0, 1)$. The training objective of our proposed method is to approximate the continuous vector field by minimizing the distance between each data point and its prediction, i.e. video frame x_t and the generated image $\hat{x}_t$.

B. Text-image-to-Video ODE

The overall architecture of the proposed method, TiV-ODE, is illustrated in Figure 2. Our approach uses the VQ-VAE [37] model for image generation. Given the initial static image, x_0, and the text caption, s, the input image x_0 is encoded as a set of image embeddings by the VQ-VAE encoder, while the text caption s is tokenized and encoded into a set of text embeddings using BERT [38]. After that, the image embeddings and text embeddings are aligned and fused using a multi-modal transformer [39], [40]. Image embeddings are used as **Query**, while text embeddings are used as **Key** and **Value**. The Text-image embeddings generated by the transformer are then used as the initial condition of the Neural ODE [19]. Afterward, using this initial condition, the Neural ODE module learns the underlying dynamical system behind the videos by approximating the continuous vector field during the training phase. Hence, the latent vector for any time point t can be generated by solving the Neural ODE at time t. The generated latent vector is then quantized by the codebook and decoded by the VQ-VAE decoder to generate a video frame $\hat{x_t}$ at time t.

C. VQ-VAE for Image Generation

The VQ-VAE-based encode-decoder structure [37] is used in our proposed method for image generation. It is important to note that before training our TiV-ODE, the VQ-VAE [37] module is pre-trained separately on each dataset and then fine-tuned to make the codebook more suitable for representing the video frames. A typical VQ-VAE model is composed of an encoder $\mathcal{E}$, a decoder $\mathcal{D}$, and a discrete codebook $\mathcal{Q} \in \mathbb{R}^{K \times N}$, which is basically a list of vectors $\{e_0, e_1, \ldots, e_K\}$, where K is the size of the codebook and N is the dimension of the codebook. The encoder encodes input image $x \in \mathbb{R}^{H \times W \times 3}$ into a latent vector $z_e(x) = \mathcal{E}(x) \in \mathbb{R}^{h \times w \times c}$,

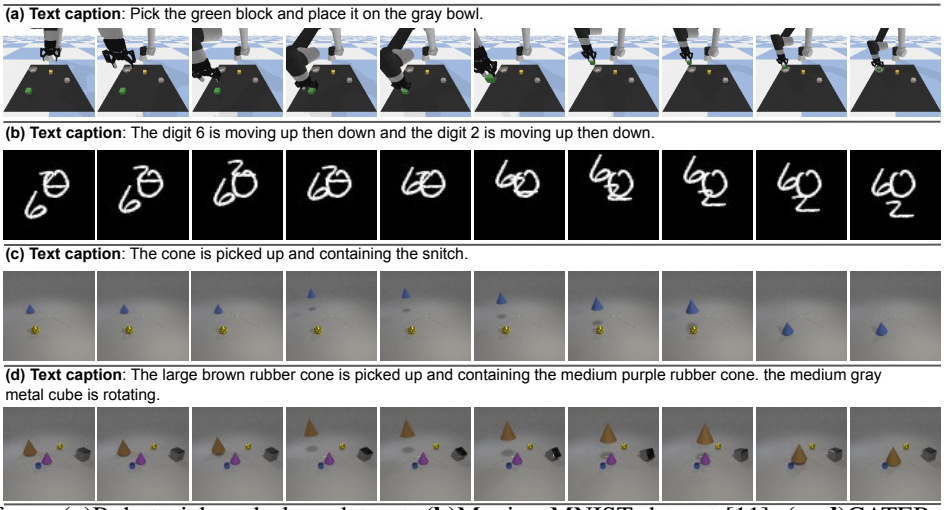

Fig. 3: Samples from: **(a)**Robot pick-and-place dataset; **(b)**Moving MNIST dataset [11]; **(c, d)**CATER datasets [35], [36].

where $h = H/n, w = W/n$, n is the downsampling ratio of the encoder, and c is the output dimension of the encoder. Then the latent vector $z_e(x)$ is compared to all vectors in the codebook, and the closest codebook vector (Euclidean distance) is input into the decoder to generate the reconstructed image. Mathematically, this is written as $\hat{x} = \mathcal{D}(z_q(x))$, where $z_q(x) = e_k$, $k = \text{argmin}_i \|z_e(x) - e_i\|_2$. The training objective of VQ-VAE is to minimize:

$$\log(p(x|z_q(x))) + \|\text{sg}[z_e(x)] - e\|_2^2 + \beta\|z_e(x) - \text{sg}[e]\|_2^2, \quad (1)$$

where $\text{sg}[*]$ stands for the stop gradient operation. The first term is the standard reconstruction loss. The second and third terms are the codebook alignment loss to make the selected codebook vector e_k close to the latent vector $z_e(x)$ by updating the codebook and encoder respectively, β is the commit loss weight. The VQ-VAE in our method is trained using the Expectation Maximization (EM) algorithm [41].

D. Text-Image Fusion Module

Inspired by the MLIM proposed in [40], a multi-modal transformer [39] is used in our TiV-ODE to fuse the input image and the text caption. Specifically, the image embedder is the encoder of a pre-trained VQ-VAE. The 2D positional embedding, similar to the one in [39], [42], is added to each image token to keep the positional information. The text caption is firstly tokenized by the BERT's [38] tokenizer, then, the text embeddings are obtained from BERT's [38] word embeddings. The positional embeddings for text tokens come with the BERT's [38] word embeddings. During the multi-modal transformer operation, the image embeddings are used as **Query**, while the text embeddings are used as **Key** and **Value**.

E. Neural ODE for Modeling Dynamical System

Neural ODE is the essential part of our proposed TiV-ODE, which models the underlying dynamical system in the latent space as a continuous ordinary differential equation. To better represent the complex dynamical systems (e.g. the trajectories of moving objects are overlapped with each other), we adopted the augmented Neural ODE [43] instead of the original one [19].

Let ξ_t be the state of the dynamical system in the latent space at an arbitrary time t, and let f be the ordinary differential equation that describes the dynamical system. The differential function f is approximated by an estimator $f_\theta \simeq f$ parameterized by θ. A time-dependent convolutional network is used in our method as the f_θ. Then, the dynamical system modeled by the Neural ODE satisfies a Cauchy problem, $\frac{\partial \xi(t)}{\partial t} = f_\theta(\xi(t), t)$, where $\xi_0 = \text{Transformer}(z_e(0), s)$. Thus, the state of the dynamical system can be obtained at any timestep by invoking an ODE solver (e.g. Runge-Kutta of Dormand-Prince [44] in our setting) to compute a numerical approximation of the integral of the dynamical system from the initial value:

$$\begin{aligned} \xi(t_i) &= \text{ODESolver}(f_\theta, \xi_0, (t_0, t_i)) \\ &\simeq \xi_0 + \int_{t_0}^{t_i} f(\xi(\tau), s, \tau) d\tau = \xi_i. \end{aligned} \quad (2)$$

Then, the state of the dynamical system in latent space at timestep t, $\xi(t)$, is quantized by the codebook $\mathcal{Q}$ and decoded by the VQ-VAE decoder $\mathcal{D}$ to reconstruct the video frame at timestep t, $\hat{x}_t$.

IV. Experiments

In this section, we first introduce the datasets used to evaluate our method. Samples from each dataset are depicted in Figure 3. Then we present the quantitive results on these datasets and compare our method with MAGE [20]. After that, we demonstrate the controllability of video generation (See Section IV-C) and the ability to model continuous dynamical systems (See Section IV-D) of our method by presenting videos generated with different image-text pairs, and videos with different frame rates respectively. Finally, our ablation studies validated the effectiveness of our model designs. For a better understanding of the videos generated by our model, we highly recommend checking the **supplementary video** accompanying this paper. Moreover, the training and implementation details of our TiV-ODE are presented in the support material. The video demo of our TiV-ODE is available at `https://youtu.be/2nQKKcgLZ28`.

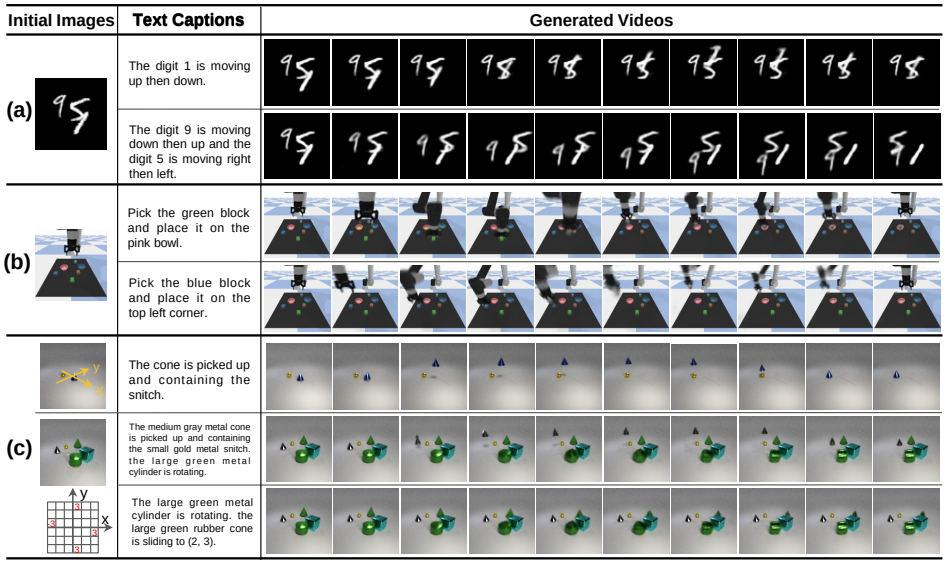

Fig. 4: Results of controllable video generation on **(a)** robot moving MNIST dataset [11], **(b)** synthetic pick-and-place dataset, and **(c)** CATER datasets [35], [36]. The coordinate system used in CATER datasets is demonstrated in (c). It is noteworthy that given the fixed initial image, our TiV-ODE can precisely manipulate different objects specified by different text captions and generate videos with both visual consistency and motion consistency.

A. Datasets

Modified Moving MNIST dataset [11], [20]. Instead of the original moving MNIST datasets [11], we used a modified version introduced in [20]. Five motion patterns are included in moving MNIST datasets, up then down, left then right, down then up, right then left, and static. We use three types of moving MNIST datasets to evaluate our method: single digits, double digits, and triple digits.

CATER datasets [35], [36] were introduced in [36] based on the CLEVR dataset [35]. There are four different motion patterns in the dataset, "contain", "slide", "rotate", and "pick-place". Each video in the dataset contains one or two random actions. We follow the same settings used in [20] to generate CATER-v1 dataset and CATER-v2 dataset. The CATER-v1 dataset contains scenes with 2 objects and one random motion. The CATER-v2 dataset contains scenes with 3 to 6 objects with two random motions.

Synthetic Robot Pick-and-Place dataset. We propose the synthetic robot pick-and-place dataset based on the simulation environment used in [45], [46]. Each sample in this dataset contains a video sequence depicting a robot pick-and-place process and a text caption specifying the pick-up and placement targets. We constructed this dataset and used it to evaluate our model, showing that our method is capable of generating videos depicting intricate robotics processes. Our results highlight the potential of our model for future robotics research.

B. Quantitive Results

Quantitative results of our TiV-ODE and comparisons against MAGE on the datasets mentioned in Section IV-A are presented in Table I, II, III. Both conventional pixel-base metrics (SSIM [47]) and perceptual metrics (image-level Fréchet inception distance (FID) [48], and learned perceptual image patch similarity (LPIPS) [49]) are used as the evaluation metrics. It is important to note that, to the best of our knowledge, there exist no metrics for evaluating the success rate of video generation w.r.t. image-text guidance. The SSIM and PSNR results of MAGE are directly taken from their paper while the FID and LPIPS results of MAGE are reproduced by using the implementations from TorchMetrics [50]. The results show that our method outperforms MAGE in terms of FID, LPIPS, while performing competi-

TABLE I: Quantitive results on moving MNIST datasets [11]. The quantitative results of MAGE are taken from [20].

Datasets	Methods	SSIM ↑	PSNR ↑
Single moving MNIST	MAGE [20]	0.97	**33.89**
	TiV-ODE (Ours)	0.97	31.8
Double moving MNIST	MAGE [20]	0.87	**24.66**
	TiV-ODE (Ours)	**0.90**	23.95
Modified double moving MNIST	MAGE [20]	0.85	**23.24**
	TiV-ODE (Ours)	0.85	21.41

TABLE II: Quantitive results on CATER-v1 dataset and CATER-v2 dataset [11]. The quantitative results of MAGE are produced using the official implementation plus our metrics implementation.

Datasets	Methods	SSIM ↑	FID ↓	LPIPS ↓	Inference speed ↓
CATER-GEN-v1	MAGE [20]	0.96	19.97	0.20	0.8s
	TiV-ODE (Ours)	0.96	**11.98**	**0.12**	**0.06s**
CATER-GEN-v2	MAGE [20]	**0.95**	**33.86**	0.20	0.8s
	TiV-ODE (Ours)	0.93	38.12	**0.18**	**0.06s**

TABLE III: Quantitive results on synthetic robot pick-and-place dataset [11]. The quantitative results of MAGE are produced using the official implementation plus our metrics implementation.

Datasets	Methods	SSIM ↑	FID ↓	LPIPS ↓
Robot Pick-and-Place dataset	MAGE [20]	**0.94**	33.69	0.18
	TiV-ODE (Ours)	0.93	**27.48**	**0.12**

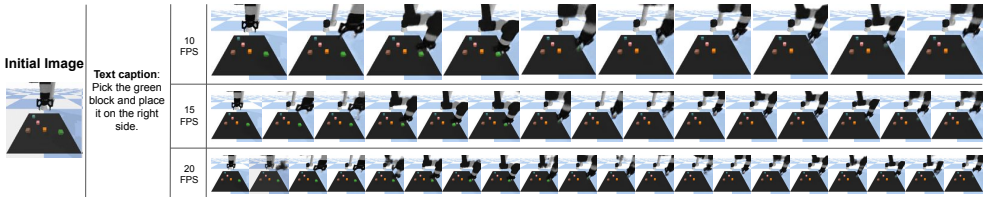

Fig. 5: Generated videos with arbitrary frame rates by solving Neural ODE with different time intervals. Each row shows the snapshots of videos of 10 FPS, 15 FPS, and 20 FPS respectively. It is noteworthy that the same video sequence is generated, but more details were presented as the frame rate increases.

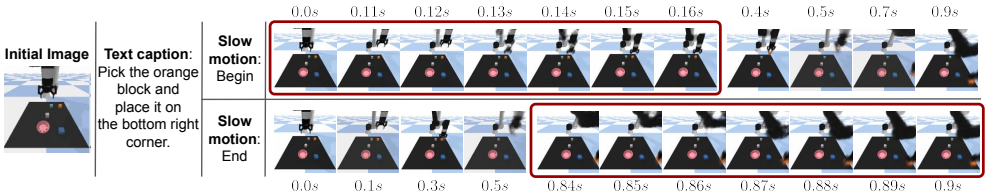

Fig. 6: Generated videos with additionally created slow-motion effects which are generated by solving the Neural ODE using a set of denser timesteps at the desired slow-motion segment. The slow-motion frames in the figure are highlighted by the red blocks, which capture more detailed slow changes compared to the normal speed segments.

tively in terms of SSIM, PSNR. The lower performance on PSNR is due to the approximation nature of the numerical ODE solvers, which may lead to the residual shadows in the generated video frames (See the digit 1 and 5 in Figure 4 (a)).

C. Controllable Video Generation

In this section, we present the test results on the moving MNIST dataset, synthetic robot pick-and-place dataset, and CATER datasets. The generated video sequences are shown in Figure 4. For each dataset, we present video sequences generated by using the same initial image but with different text captions. These results demonstrate that our model successfully learns the alignments between images and texts as well as the underlying dynamical system, making it able to generate videos with visual consistency and motion consistency from different image-text pairs. Overall, these results show that our method yields promising performance in achieving highly controllable video generation with a given static image and a text caption.

D. Video Generation with Different Frame Rates

In this section, we demonstrate the ability of our method to model the underlying continuous dynamical system by showing : (*i*) video generation with arbitrary frame rates; (*ii*) video generation with manually added slow motion effect.

Video generation with arbitrary frame rates. Our model is able to generate video sequences with arbitrary frame rates by solving the learned Neural ODE with different time intervals. Here we present the results from the synthetic robot pick-and-place dataset. Three video sequences with 10, 15, and 20 FPS are generated (See Figure 5).

Video generation with slow-motion effects. Our model is able to generate video sequences with manually added slow-motion effects that can be adjusted by using denser timesteps at the desired slow-motion segment. Formally, this effect is referred to as frame rate ramping. Here we present two generated video sequences from the synthetic robot pick-and-place dataset. One has slow motion at the beginning and the other has it at the end (See Figure 6).

TABLE IV: TiV-ODE ablation study on CATER-v1 Dataset.

Method	SSIM↑	FID↓	LPIPS↓
TiV-ODE	**0.96**	**11.98**	**0.12**
- *w irregular timesteps*	0.96	13.71	0.13
- *w/o ODE solver*	0.90	31.56	0.28

These results show that our model is capable of modeling the underlying continuous dynamical system from videos, and with the learned continuous dynamical system, our model is able to control the framerates of the generated videos, which cannot be done by previous methods [20].

E. Ablation Study

We have conducted an ablation study on the CATER-v1 dataset to justify the necessity of the Neural ODE module and the robustness of our method against irregular videos: (*i*) Videos with irregular timesteps are used to train our model. (*ii*) The Neural ODE module is replaced with a step-wise transition module, i.e. the transition model receives the video frame at t and predicts the frame at $t+1$. Quantitative results of the ablation study are presented in Table V. These results show that our method is robust to irregular videos, which can not be done by previous methods [20], and also demonstrate the effectiveness of the Neural ODE module in our proposed TiV-ODE.

V. Conclusion

This paper presents a novel controllable video generation method that generates highly controllable videos conditioned on an image-text pair. Moreover, our framework models and learns the underlying continuous dynamical system using Neural ODE. To show the potential of our model in robotics research, we created a new robot pick-and-place dataset for evaluation, as well as using the existing moving MNIST datasets and CATER datasets. Experiment results showed that our method yields promising results in terms of controllable video generation and dynamical system modeling. This work moves a significant step towards solving the challenging controllable video generation task and has the potential for downstream applications in robotics.

TABLE V: Comparison with SOTA methods.

Method	PSNR↑	SSIM↑	LPIPS↓
OCVT	31.08	0.88	0.13
Slot-LSTM	32.15	0.90	0.09
Slotformer	32.04	0.91	0.08
Ours			
- *w/o residual*	**39.96**	**0.99**	**0.01**
- *w residual*	**41.21**	**0.99**	**0.01**

REFERENCES

[1] C. Choi, J. H. Choi, J. Li, and S. Malla, "Shared cross-modal trajectory prediction for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 244–253.

[2] Y. Chen, F. Rong, S. Duggal, S. Wang, X. Yan, S. Manivasagam, S. Xue, E. Yumer, and R. Urtasun, "Geosim: Realistic video simulation via geometry-aware composition for self-driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7230–7240.

[3] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.

[4] Y. Dai, M. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," *arXiv preprint arXiv:2302.00111*, 2023.

[5] I. Kapelyukh, V. Vosylius, and E. Johns, "Dall-e-bot: Introducing web-scale diffusion models to robotics," *arXiv preprint arXiv:2210.02438*, 2022.

[6] M. Attarian, A. Gupta, Z. Zhou, W. Yu, I. Gilitschenski, and A. Garg, "See, plan, predict: Language-guided cognitive planning with video prediction," *arXiv preprint arXiv:2210.03825*, 2022.

[7] Z. Hao, X. Huang, and S. Belongie, "Controllable video generation with sparse trajectories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7854–7863.

[8] J. Pan, C. Wang, X. Jia, J. Shao, L. Sheng, J. Yan, and X. Wang, "Video generation from single semantic label map," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3733–3742.

[9] L. Sheng, J. Pan, J. Guo, J. Shao, and C. C. Loy, "High-quality video generation from static structural annotations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2552–2569, 2020.

[10] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, "Videogpt: Video generation using vq-vae and transformers," *arXiv preprint arXiv:2104.10157*, 2021.

[11] G. Mittal, T. Marwah, and V. N. Balasubramanian, "Sync-draw: Automatic video generation using deep recurrent attentive architectures," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1096–1104.

[12] C. Wu, L. Huang, Q. Zhang, B. Li, L. Ji, F. Yang, G. Sapiro, and N. Duan, "Godiva: Generating open-domain videos from natural descriptions," *arXiv preprint arXiv:2104.14806*, 2021.

[13] S. Park, K. Kim, J. Lee, J. Choo, J. Lee, S. Kim, and E. Choi, "Vid-ode: Continuous-time video generation with neural ordinary differential equation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2412–2422.

[14] A. Blattmann, T. Milbich, M. Dorkenwald, and B. Ommer, "Understanding object dynamics for interactive image-to-video synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5171–5181.

[15] M. Jin, Z. Hu, and P. Favaro, "Learning to extract flawless slow motion from blurry videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8112–8121.

[16] A. Paliwal and N. K. Kalantari, "Deep slow motion video reconstruction with hybrid imaging system," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 7, pp. 1557–1569, 2020.

[17] B. Chang, M. Chen, E. Haber, and E. H. Chi, "Antisymmetricrnn: A dynamical system view on recurrent neural networks," *arXiv preprint arXiv:1902.09689*, 2019.

[18] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[19] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[20] Y. Hu, C. Luo, and Z. Chen, "Make it move: controllable image-to-video generation with text descriptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 219–18 228.

[21] Z. Lin, Y.-F. Wu, S. Peri, B. Fu, J. Jiang, and S. Ahn, "Improving generative imagination in object-centric world models," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6140–6149.

[22] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, and K. Greff, "Conditional object-centric learning from video," *arXiv preprint arXiv:2111.12594*, 2021.

[23] N. Li, M. A. Raza, W. Hu, Z. Sun, and R. Fisher, "Object-centric representation learning with generative spatial-temporal factorization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 772–10 783, 2021.

[24] Z. Wu, N. Dvornik, K. Greff, T. Kipf, and A. Garg, "Slotformer: Unsupervised visual dynamics simulation with object-centric models," *arXiv preprint arXiv:2210.05861*, 2022.

[25] J. Xie, S.-C. Zhu, and Y. Nian Wu, "Synthesizing dynamic patterns by spatial-temporal generative convnet," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 7093–7101.

[26] J. Xie, S.-C. Zhu, and Y. N. Wu, "Learning energy-based spatial-temporal generative convnets for dynamic patterns," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 516–531, 2019.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] Y. Rubanova, R. T. Chen, and D. K. Duvenaud, "Latent ordinary differential equations for irregularly-sampled time series," *Advances in neural information processing systems*, vol. 32, 2019.

[29] E. De Brouwer, J. Simm, A. Arany, and Y. Moreau, "Gru-ode-bayes: Continuous modeling of sporadically-observed time series," *Advances in neural information processing systems*, vol. 32, 2019.

[30] C. Yildiz, M. Heinonen, and H. Lahdesmaki, "Ode2vae: Deep generative second order odes with bayesian neural networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[31] T. Teshima, K. Tojo, M. Ikeda, I. Ishikawa, and K. Oono, "Universal approximation property of neural ordinary differential equations," *arXiv preprint arXiv:2012.02414*, 2020.

[32] H. Zhang, X. Gao, J. Unterman, and T. Arodz, "Approximation capabilities of neural odes and invertible residual networks," in *International Conference on Machine Learning*. PMLR, 2020, pp. 11 086–11 095.

[33] D. Kanaa, V. Voleti, S. E. Kahou, and C. Pal, "Simple video generation using neural odes," *arXiv preprint arXiv:2109.03292*, 2021.

[34] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[35] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[36] R. Girdhar and D. Ramanan, "Cater: A diagnostic dataset for compositional actions and temporal reasoning," *arXiv preprint arXiv:1910.04744*, 2019.

[37] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[40] T. Arici, M. S. Seyfioglu, T. Neiman, Y. Xu, S. Train, T. Chilimbi, B. Zeng, and I. Tutar, "Mlim: Vision-and-language model pre-training with masked language and image modeling," *arXiv preprint arXiv:2109.12178*, 2021.

[41] A. Roy, A. Vaswani, A. Neelakantan, and N. Parmar, "Theory and experiments on vector quantized autoencoders," *arXiv preprint arXiv:1805.11063*, 2018.
[42] Z. Wang and J.-C. Liu, "Translating math formula images to latex sequences using deep neural networks with sequence-level training," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, no. 1-2, pp. 63–75, 2021.
[43] E. Dupont, A. Doucet, and Y. W. Teh, "Augmented neural odes," *Advances in neural information processing systems*, vol. 32, 2019.
[44] J. R. Dormand and P. J. Prince, "A family of embedded runge-kutta formulae," *Journal of computational and applied mathematics*, vol. 6, no. 1, pp. 19–26, 1980.
[45] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv*, 2022.
[46] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022.
[47] Z. 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*, vol. 13, no. 4, pp. 600–612, 2004.
[48] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
[49] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," *Advances in neural information processing systems*, vol. 29, 2016.
[50] N. S. Detlefsen, J. Borovec, J. Schock, A. H. Jha, T. Koker, L. D. Liello, D. Stancl, C. Quan, M. Grechkin, and W. Falcon, "Torchmetrics - measuring reproducibility in pytorch," *Journal of Open Source Software*, vol. 7, no. 70, p. 4101, 2022. [Online]. Available: https://doi.org/10.21105/joss.04101