Make It Move: Controllable Image-to-Video Generation with Text Descriptions

Yaosi Hu[1*] Chong Luo[2] Zhenzhong Chen[1]
Wuhan University[1] Microsoft Research Asia[2]
ys_hu@whu.edu.cn cluo@microsoft.com zzchen@whu.edu.cn

Abstract

Generating controllable videos conforming to user intentions is an appealing yet challenging topic in computer vision. To enable maneuverable control in line with user intentions, a novel video generation task, named Text-Image-to-Video generation (TI2V), is proposed. With both controllable appearance and motion, TI2V aims at generating videos from a static image and a text description. The key challenges of TI2V task lie both in aligning appearance and motion from different modalities, and in handling uncertainty in text descriptions. To address these challenges, we propose a Motion Anchor-based video GEnerator (MAGE) with an innovative motion anchor (MA) structure to store appearance-motion aligned representation. To model the uncertainty and increase the diversity, it further allows the injection of explicit condition and implicit randomness. Through three-dimensional axial transformers, MA is interacted with given image to generate next frames recursively with satisfying controllability and diversity. Accompanying the new task, we build two new video-text paired datasets based on MNIST and CATER for evaluation. Experiments conducted on these datasets verify the effectiveness of MAGE and show appealing potentials of TI2V task. Datasets are available at https://github.com/Youncy-Hu/MAGE.

1. Introduction

Video generation has undergone revolutionary changes and has made great progress in recent years. Early research of unconditional video generation [24, 26, 28] focused on how to generate a video from noise or a latent vector from an aligned latent space. Recently, more emphases have been put on controllable video generation [4, 11, 30], which allows users to express their intentions about how the scene or the objects look like (appearance information) or how the objects move (motion information). Controllable video generation has many potential applications, including facilitating designers in artistic creation and assisting machine learning practitioners for data augmentation.

[*]This work was done while Yaosi Hu was an intern at MSRA.

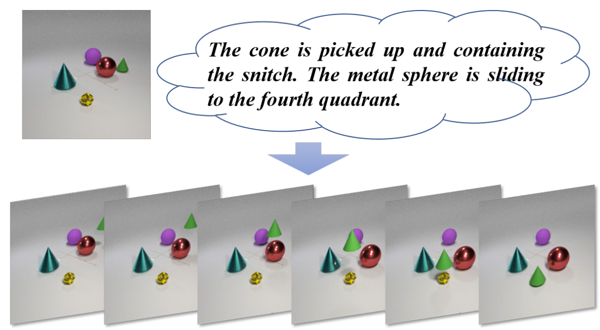


Figure 1. An illustration of the proposed TI2V task. An image and a detailed text description provide the appearance and motion information for video generation, respectively.

Existing controllable video generation tasks can be grouped into three categories, namely Image-to-Video generation (I2V), Video-to-Video generation (V2V), and Text-to-Video generation (T2V). These tasks provide different ways for users to inject the appearance and the motion information and therefore have different levels of control over these two factors. I2V and V2V have strong control over the appearance of generated video, as separate images are usually provided to set the scene. As for the motion, I2V shows limited controllability since the task is defined to accept only coarse-grained motion clues, such as predefined action labels or directions [4]. In contrast, V2V can generate videos with highly controllable motion because detailed motion guidance, such as trajectories [11] or action sequences [6, 19], are provided in the form of input video. But one drawback of V2V in practical use is that such motion guidance is hard to be obtained. Among all the three tasks, T2V has the weakest control over the generated video. Users provide both appearance and motion information through text, which is imprecise and sometimes ambiguous. Nevertheless, text description of motion is more in line with human habits [35] and leaves a lot of room for creation and imagination in video generation.

In this paper, we introduce a novel video generation task, named Text-Image-to-Video generation (TI2V). It provides a natural way for users to express their intentions, using a single static image to set the scene and a natural text description to provide motion. TI2V is a more difficult task than I2V or T2V. It not only requires the separate understanding of text and image, but also needs to align visual objects with corresponding text descriptions, and then transform the implied object motion to an explicit video. We aim to achieve two goals in the TI2V task: i) **Controllable**. Under the constraints of image and text, the generated video should have visually consistent appearance set by the given image and semantically aligned motion as described in the text. ii) **Diverse**. This goal resolves ambiguity and brings creativity, which are important and appealing features for video generation. In the example given in Fig.1, the text description does not specify which cone it wants to pick up and which exact position in "the fourth quadrant" it wants the metal sphere to slide to. Under such "constrained randomness", we want to produce videos that match the description but are also diverse.

We design an auto-regressive framework, named MAGE, to address the TI2V task. A VQ-VAE encoder-decoder architecture is adopted for efficient visual token representation. The key challenge is how to merge the text-described motion into visual features to generate a controllable and diverse video. To achieve the controllable goal, we propose a spatially aligned Motion Anchor (MA) to integrate the appearance and motion information through the cross-attention operation in the common latent space for image and text. Each position in the MA stores all necessary motion information of the corresponding region for video generation. We further introduce explicit condition and implicit randomness into MA. The explicit condition provides additional constraint from an explicit input (e.g., speed) to improve both controllability and diversity, while the implicit randomness brings in uncertainty in the data distribution, allowing the model to generate diverse videos in a stochastic way. In the proposed MAGE framework, we adopt axial transformer to inject and fuse MA into visual tokens and generate videos in an auto-regressive manner.

To evaluate TI2V task and our generation model, appropriate paired video-text datasets are in need. Different from T2V task that often conducts experiments on action recognition datasets like KTH [25] or captioning datasets like MSR-VTT [38] with action label or coarse-grained caption, TI2V focuses more on the maneuvering capability to image and requires fine-grained text description. Therefore, we propose two datasets with synthetic videos and fine-grained text descriptions based on MNIST [14] and CATER [10] for TI2V task. By controlling the uncertainty in descriptions, we can evaluate the performance of both deterministic and diverse video generation.

The contributions of this paper are concluded as follows:

- A novel Text-Image-to-Video generation task (TI2V) is introduced, aiming to generate visually consistent video from an image and a text description.
- A Motion Anchor-based video GEnerator (MAGE) is proposed to generate controllable and diverse videos. The core structure, motion anchor (MA), addresses the challenging matching problem between the appearance in the image and the motion clues in the text.
- Two video-text paired datasets modified from MNIST and CATER are built for the evaluation of TI2V task. Moreover, experiments conducted on these two datasets verify the effectiveness of MAGE.

2. Related Work

We only consider the work of video generation guided by human intention. In this context, human intention is mainly composed of the description of the scene (spatial information) and the description of the motion (temporal information). We classify related work into three categories, namely video-to-video (V2V) generation, image-to-video (I2V) generation, and text-to-video (T2V) generation, according to how human intention is expressed.

Note that unconditional video generation [24, 26, 28, 32, 33] is not discussed here, as these works generate videos from a random variable or a latent vector, and do not provide an interface for human to express their intention.

2.1. Video-to-Video Generation

There are two popular forms of V2V task, namely future video prediction and video-to-video synthesis. They retrieve or predict the desired motion from an input video.

Future video prediction predicts future frames based on several past frames provided. The generator is required to retrieve past motion and predict the future [17, 29, 34, 36]. Due to the unpredictable nature of object motion, works in this setting are only used to predict very few future frames.

In video-to-video synthesis, motion information is provided in an explicit form, such as a set of sparse motion trajectories [11], a sequence of human poses [6,31], or a sequence of discrete actions [9, 19]. The spatial information is provided with separate images or frames (with or without segmentation masks) [11, 19, 30] or structured data such as 3D face mesh [41]. Thanks to the rich input information, video-to-video synthesis can generate high-quality videos with controllable motions. However, the motion sequence is usually hard to obtain.

2.2. Image-to-Video Generation

Image-to-video generation refers to the line of work which generates video from a single image and a random or very coarse motion clue.

When the motion clue is not provided at all, videos are generated in a stochastic manner constrained by the spatial information provided by the input image [2, 5, 8, 15, 37, 39, 40]. The models used to generate videos can be generative adversarial network (GAN) [40] or variational autoencoder (VAE) [8]. This kind of stochastic video generation can only handle short dynamic patterns in distribution.

In order to produce more controllable video, coarse-grained motion clues, including predefined directions or action labels, can be provided [21]. Recently, Blattmann *et al.* [4] propose an interactive I2V synthesis model which allows users to specify the desired motion through the manual poking of a pixel. I2V generation does not require users to provide detailed motion information, which reduces the burden for use, but at the same time it is unlikely to be used for generating videos with complex motion patterns.

2.3. Text-to-Video Generation

T2V task aims at generating videos just from text descriptions, which is a challenging task. There is relatively little research on this topic.

Mittal *et al.* [20] first introduced this task and proposed a VAE-based framework, called Sync-DRAW, to encode simple captions and generate semantically consistent videos. A concurrent work [18] performs variable-length semantic video generation from captions. The model relies on VAE and recurrent neural networks (RNNs) to learn the long-term and short-term context of the video. Other works *et al.* [3, 7, 16, 22] also try to generate video from the caption and a latent noise vector. However, due to the ambiguity of text and its inefficacy in providing fine-grained appearance information, the generated videos are usually at low resolution or very blurry. Recently, a VQ-VAE-based generator named GODIVA [35] was proposed to generate open-domain videos from text. However, the appearance of the generated video tends to be the most salient feature that has been seen in the training phrase. It is unlikely to generate videos for complex scene or unseen appearance.

Compared to I2V, the proposed TI2V can generate controllable video through maneuverable dynamics. Compared to V2V, TI2V does not need complex auxiliary information. Compared to T2V, TI2V can generate more controllable video due to the specified appearance and fine-grained motion.

3. MAGE for Text-Image-to-Video Task

3.1. TI2V Problem Formulation

TI2V task aims at generating a video from a static image and a text description. Formally, given a single static image $x_1 \in \mathbb{R}^{h \times w \times C}$ and a text description $s = \{s_1, \cdots, s_L\}$ containing L words, the goal of TI2V is to learn a mapping function that generates a sequence of frames $\hat{x} = \{\hat{x}_2, \cdots, \hat{x}_T\}$ with consistent appearance with x_1 and specified motion in s.

We address the TI2V task through a supervised learning approach. During training, a reference video, denoted by $x = \{x_2, \cdots, x_T\}$ is provided for (x_1, s). The training objective is to make the conditional distribution of $\hat{x}$ given x_1 and s approximate the conditional distribution of x.

3.2. MAGE Framework Overview

MAGE adopts a VQ-VAE-based encoder-decoder architecture. VQ-VAE [27] is an effective tool to convert highly redundant visual data into a concise representation, which is a good choice for semantic-level manipulation, as we need to do in video generation.

The entire framework is illustrated in Fig. 2. When an initial image x_1 and a text description s is given, the image is passed to the VQ-VAE encoder and tokenized into a group (16x16) of latent codes z_1. The vector quantized image token, as well as text embeddings are then passed to the cross-attention module to obtain a spatially aligned motion representation, called motion anchor ($\widetilde{M}$). The explicit condition and implicit randomness, which will be detailed later, are also integrated into the MA.

Then, the MA $\widetilde{M}$ is fused with z_1 by axial transformer blocks to produce $\hat{z}_2$, from which the VQ-VAE decoder can decode the next video frame $\hat{x}_2$. Once $\hat{z}_i$ ($i \geq 2$) is obtained, it is sent back to the axial transformers to generate $\hat{z}_{i+1}$ together with $\widetilde{M}$ and all previous frames. Note that MA is a global variable which contains all the motion information needed to generate the entire sequence, so it only needs to be computed once. The generation process stops when a predefined sequence length is reached.

The core of the MAGE framework is the motion anchor. All the related network models, including the cross-attention, the axial transformer blocks, the function ϕ that encodes explicit condition, and the function ψ that encodes implicit randomness are trained together to achieve a unified objective. The VQ-VAE encoder and decoder, however, are peripheral modules which are trained beforehand.

For completeness, we first introduce how we pre-train VQ-VAE. A VQ-VAE model consists of a latent codebook $C \in \mathbb{R}^{K \times D}$, an encoder E and decoder D with respective down-sampling and up-sampling ratio n. The input image $x \in \mathbb{R}^{H \times W \times C}$ is encoded into latent vector $e_x = E(x) \in \mathbb{R}^{h \times w \times D}$ first, where $h = H/n, w = W/n$. Then e_x is discretized by a nearest neighbour look-up in C to get both quantized index $z \in \mathbb{R}^{h \times w}$ and quantized vector $\tilde{e}_x \in \mathbb{R}^{h \times w \times D}$. The decoder, with a reversed structure of the encoder, then reconstructs image $\hat{x}$ from $\tilde{e}_x$.

VQ-VAE is trained with an image-level reconstruction task. The training objective consists of the reconstruction loss, codebook loss, and commitment loss. It is written as:

$$\mathcal{L} = \log P\left(\hat{x} \mid \tilde{e}_x\right) + \left\|\mathrm{sg}\left(e_x\right) - \tilde{e}_x\right\|_2^2 + \beta \left\|e_x - sg\left(\tilde{e}_x\right)\right\|_2^2, \tag{1}$$

where sg stands for the stop-gradient operator and β is the weighting factor. After this pre-training stage, parameters

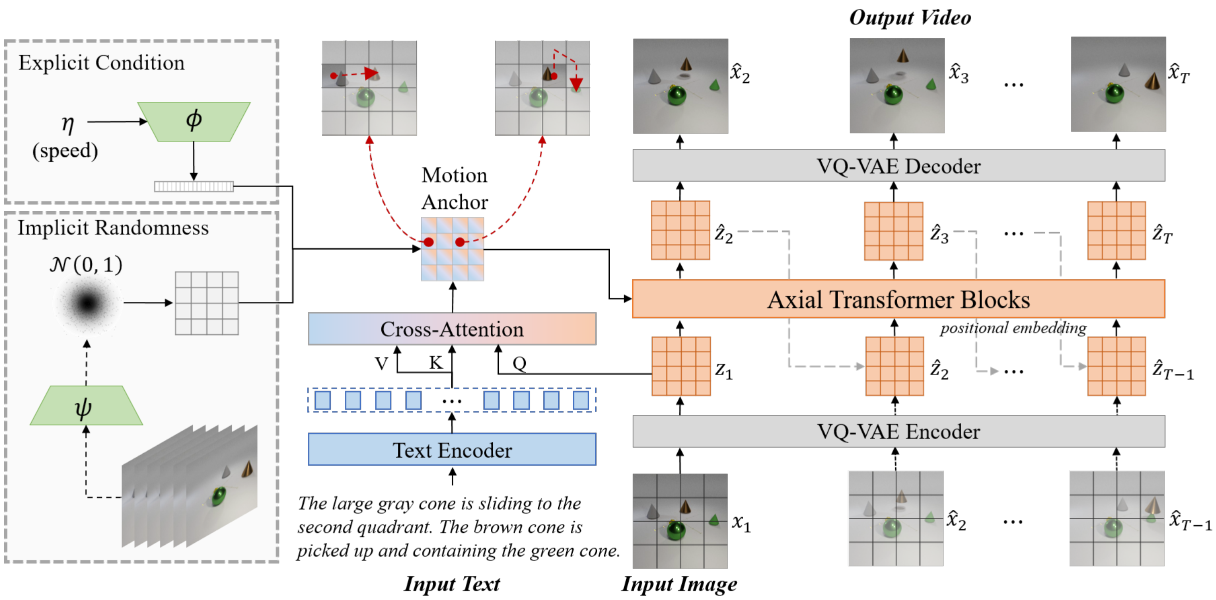


Figure 2. Illustration of the MAGE framework. The black dashed lines and grey dashed lines stand for operations that are only involved in training and inference processes, respectively. The black solid lines indicate the operations used in both processes. Each frame is represented by 4×4 tokens in the figure only for visualization purpose.

in E and D are frozen.

3.3. MA-Based Video Generation

Motion anchor is the core design in MAGE to achieve controllable and diverse video generation for the TI2V task. It aligns text with image, and allows for the injection of explicit condition and implicit randomness.

3.3.1 Image-Text Alignment

In MAGE, we employ a cross-attention module to achieve the alignment between the image content and the motion clues given in text. We employ a learnable text encoder to compute the text embedding $e_s \in \mathbb{R}^{L\times d}$ from the input text s, where d is the hidden size. e_s is used as the key and the value of the cross-attention module. The image embedding z_1 is converted to the same latent space as e_s by a learnable embedding matrix, and the converted embedding, denoted by $e_{z_1} \in \mathbb{R}^{h\times w\times d}$, is used as the query.

The cross-attention operation locates the responsive words for each visual token and aggregates the implied motion information. Then, the motion information and the corresponding visual information are fused by a feed forward and normalization layer to generate the motion anchor $M \in \mathbb{R}^{h\times w\times d}$. This process can be described by:

$$\begin{aligned} &Q = e_{z_{1,i,j}}W^q, K = e_sW^k, V = e_sW^v, \\ &A_{i,j} = MultiHead(Q,K,V), \\ &M_{i,j} = FFN\left(e_{z_{1,i,j}}, A_{i,j}\right). \end{aligned} \tag{2}$$

where $e_{z_{1,i,j}}$ stands for the visual embedding at position (i,j) in the first frame. And $MultiHead$ and FFN represent multi-head attention and feed forward network, respectively. Thus, $M_{i,j}$ stores the appearance and motion information at position (i,j).

3.3.2 Explicit Condition and Implicit Randomness

We allow MA to encode some quantifiable conditions that are not expressed in the text. In this work, we demonstrate how a typical condition known as motion speed is incorporated into the generation process and reflected in the generated video. A simple linear layer ϕ is applied to encode speed η into an embedding vector $c \in \mathbb{R}^d$. It is written as:

$$c = \phi(\eta). \tag{3}$$

Besides, text description can be ambiguous. The "correct" video that matches an input image-text pair may not be unique. Thus, the model is required to accommodate existing randomness r in data distribution and randomly generate diverse videos which are semantically consistent with the text. We propose to use a variational information bottleneck [1] ψ for implicit randomness modeling. ψ consists of several 3D convolutional blocks and a reparameterization layer. During training, we encode the video randomness into a random variable that conforms to the standard normal distribution. During inference, a random variable is directly sampled from the distribution and merged into the

MA,

$$r \sim q_{\psi}\left(r \mid e_{z_1 \sim T}\right). \tag{4}$$

To inject randomness r into MA, an adaptive instance normalization (AdaIN) layer [13] is applied. Since speed affects each movement equivalently, c is directly injected into M through and a channel-wise additive to change the global motion information. It is formulated as

$$\widetilde{M} = AdaIN(M, r) + c. \tag{5}$$

3.3.3 Appearance-Motion Fusion

After obtaining the motion anchor $\widetilde{M} \in \mathbb{R}^{h \times w \times d}$, the video generator G jointly models MA and visual token embeddings. To reduce computation, we adopt N axial transformer blocks which consist of three-dimensional axial attention [12, 35] on temporal-wise, row-wise, and column-wise, respectively. As such, the attention complexity is reduced from $\mathcal{O}\left((Thw)^2\right)$ to $\mathcal{O}\left(Thw\left(T+h+w\right)\right)$. The generation can be formulated as

$$\hat{z}_i = G\left(PE\left(\left[\widetilde{M} \cdot z_{<i}\right]\right)\right), \tag{6}$$

where $[\cdot]$ stands for the concatenation operation and PE represents positional embedding. Noted that row-wise and column-wise attention have full receptive field on respective axis. But for temporal-wise axial attention, we apply a causal mask to ensure that a visual token can only receive the information from previous frames. After stacking several axial transformer blocks, each visual token has full receptive field on current and previous frames for spatial and temporal information. A token in each position can not only get the complete motion information from the MA, but also "track" the motion in previous frames. Recurrently generating frames with spatially aligned MA ensures coherent and consistent video output.

The training objective for video generator consists of a cross-entropy loss for visual token prediction, and two constraints for explicit condition c and implicit randomness $q_{\psi}(r \mid X)$. It is formulated as

$$\begin{aligned} \mathcal{L} = &-\frac{1}{T}\sum_{i=2}^{T} z_i \log\left(P\left(\hat{z}_i \mid z_{<i}, s, c, r\right)\right) \\ &+ \alpha \|c\|_2^2 + \beta KL\left(q_{\psi}(r \mid X) \| p(r)\right), \end{aligned} \tag{7}$$

where α and β are hyper-parameters to trade-off the two constraints, and $p(r)$ stands for the standard normal prior.

4. Experiments

4.1. Datasets

Traditional T2V methods are often evaluated on Single Moving MNIST and Double Moving MNIST [20] that contain one and two moving digits, respectively. Four motion patterns are included: right then left, left then right, up then down, down then up. Once the border is reached, the digit will rebound. We propose a Modified Double Moving MNIST dataset with more movements and a distracting digit. We also propose two versions of synthetic CATER-GEN datasets in a 3D environment with lighting and shadows built upon CATER [10]. These datasets are briefly introduced as follows:

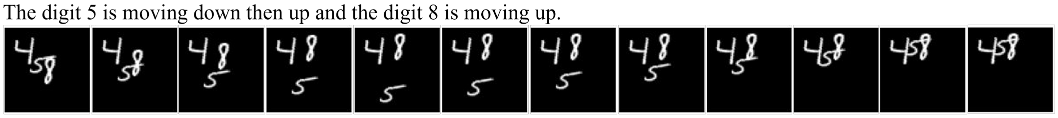
- **Modified Double Moving MNIST**: In order to control the moving process, we keep two options for the movement along each direction. One is to stop at the edge and the other is to bounce once. Once the action finishes, the digit will stay still. Thus, we get 8 motion patterns for all four directions. Besides, we randomly insert one static distracting digit other than the moving ones as background.

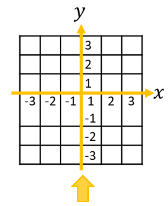
- **CATER-GEN-v1**: CATER-GEN-v1 is a simpler version which is built with two objects (cone and snitch) and a large "table" plane inherited from CATER. There exist four atomic actions: "rotate", "contain", "pick-place" and "slide". Each video randomly contains one or two actions. When generating descriptions, we design a predefined sentence template to fill the subject, action, and optional object. The final position is also provided for actions "pick-place" and "slide". By specifying the final position with an accurate coordinate or a quadrant area, explicit descriptions and ambiguous descriptions are provided for deterministic and diverse video generation, respectively.

- **CATER-GEN-v2**: CATER-GEN-v2 is a much more complex dataset which contains $3 \sim 8$ objects in each video. Each object has 4 attributes that are randomly chosen from five shapes, three sizes, nine colors, and two materials. The atomic actions are the same as in CATER-GEN-v1. To create ambiguity in text descriptions, we not only replace the final coordinate, but also randomly discard the attributes for each object, thus the object may not be unique due to the uncertainty in the referring expression.

The resolution of the generated video is 64×64 for the new and two existing MNIST-based datasets, and 256×256 for two CATER-based datasets. Both Single Moving MNIST and Double Moving MNIST contain 10k pairs for training and 2k for testing following [7]. For CATER-GEN-v1, we generate 3.5k pairs for training and 1.5k pairs for testing. For the more complicated datasets Modified-MNIST and CATER-GEN-v2, we generate 24k pairs for training and 6k pairs for testing. Samples from generated datasets are shown in Fig. 3.

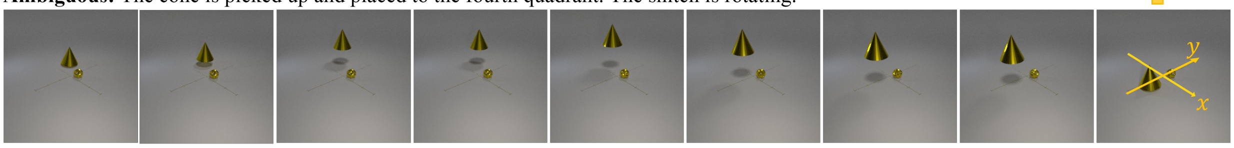
The digit 5 is moving down then up and the digit 8 is moving up.





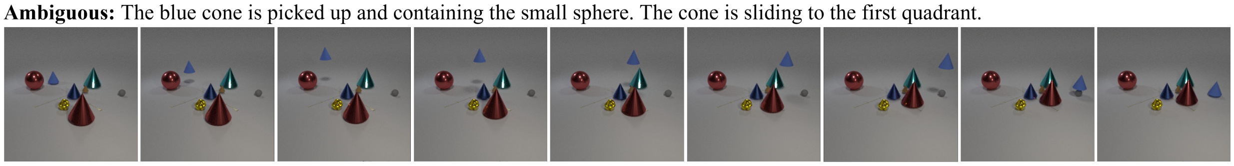
(a) Modified Double Moving MNIST

Explicit: The cone is picked up and placed to (1, -2). The snitch is rotating.
Ambiguous: The cone is picked up and placed to the fourth quadrant. The snitch is rotating.



(b) CATER-GEN-v1

Explicit: The medium blue rubber cone is picked up and containing the small gray rubber sphere. The large red metal cone is sliding to (2, 1).
Ambiguous: The blue cone is picked up and containing the small sphere. The cone is sliding to the first quadrant.



(c) CATER-GEN-v2

Figure 3. Samples from Modified Double Moving MNIST, CATER-GEN-v1 and CATER-GEN-v2, respectively. For CATER-GENs, we split the 2D table plane into a 6×6 portion with fixed axes as shown on the top right. Thus, the position of object can be described with coordinate or quadrant. Note that the camera position is static.

4.2. Implementation Details

Both VQ-VAE and video generator in our experiments are trained from scratch. For VQ-VAE, we use the similar encoder and decoder structure as [23] with codebook size 512×256. The input size is $H = 64, W = 64, C = 1$ for MNIST-based datasets, and $H = 128, W = 128, C = 3$ for CATER-based datasets. After VQ, images are compressed to 16×16 visual tokens. The text description is encoded by a two-layer transformer. We let $T = 10, d = 512$ and stack two axial transformer blocks to generate the video. We use a batch size of 32 and a learning rate of 5e-5. The speed η is normalized to $(0, 1)$, which is mapped to a predefined frame sampling interval.

In the training stage, the speed η is randomly sampled from $(0, 1)$ and we use the corresponding sampling interval to obtain the reference video. To improve training efficiency, we input s and $\{x_1, \cdots, x_{T-1}\}$ to predict future frame tokens $\{\hat{z}_2, \cdots, \hat{z}_T\}$ in parallel. For inference, only the first image and a text description are given. We sample the implicit randomness from a normal distribution and generate the video sequence in an auto-regressive way.

4.3. Deterministic Video Generation

To evaluate whether MAGE can achieve the first goal of TI2V task, known as controllability, we discard implicit randomness module and use the explicit descriptions in the dataset. Thus, the whole model is deterministic, and the "correct" video is unique. In this section, we first show the qualitative performance of generated videos (please refer to supplementary Sec.2 for quantitative results and ablation studies), then the controllability of explicit condition is evaluated. More interestingly, we try to input the same image but different descriptions (speed kept unchanged), to show how to manipulate the objects through different text.

4.3.1 Qualitative Evaluation

We first show the qualitative results on three MNIST-based datasets and two CATER-based datasets with explicit descriptions in Fig.5 and Fig.4, respectively. The generated videos have high visual quality, and the motion is highly coherent with the one specified in the text. For three MNIST-based datasets, the digits are moving smoothly and the shapes are well maintained. Especially for the hard example from Modified Double Moving MNIST (Fig.5(c)), the given image contains three digits with a large overlap. MAGE is able to recognize "3" and "9" specified in the text and disassemble them, while keeping the unmentioned "1" still.

In CATER-GEN-v1, although there are only two objects, it is still a challenge to not only generate right motion, but also simulate variations of surface lighting and shadows of objects because of three illuminants out of scene. It can be observed in Fig.4(a) that the cone is placed to the right coordinate. Both the changing of surface light and shadow are generated quite well. For the rotating of snitch, it is difficult for both VQ-VAE and the generation model to reconstruct such small object with complex appearance. We

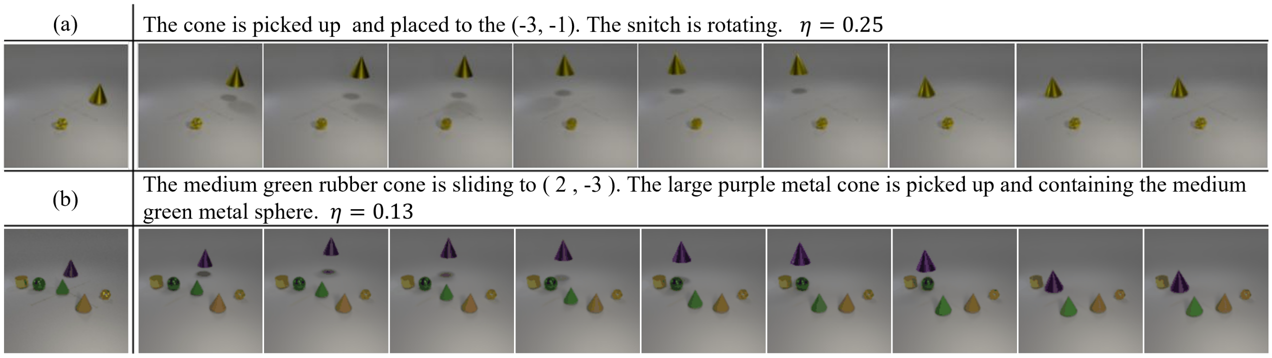


Figure 4. Generated Samples from CATER-GEN-v1 and CATER-GEN-v2, respectively. The images of the left column are given first images, and the first row of each sample is the given text and speed.

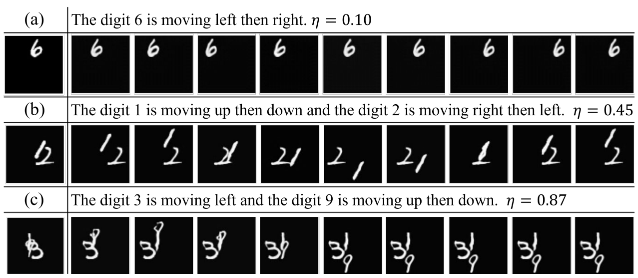


Figure 5. Generated Samples from Single / Double / Modified Double Moving MNIST, respectively, from top to bottom. The left column shows given images, and the first row of each sample is the given text and speed.

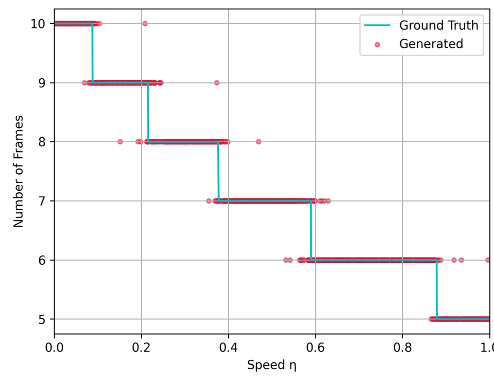


Figure 6. Statistics of the number of frames when the motion finishes from CATER-GEN-v1 test set (1500 samples), where the X-axis stands for the input speed η and Y-axis stands for the number of frames. The blue line and red scatter represent ground truth and generated video, respectively.

can still observe the variation on the surface which suggests that there is an action "rotate" happening.

For the sample from CATER-GEN-v2 as shown in Fig.4(b), although there are six different objects, the model successfully locates the "medium green rubber cone", "large purple metal cone", and "medium green metal sphere". It can also handle the occlusion relation for action "contain". Meanwhile, other unmentioned objects are kept stable and still.

4.3.2 Explicit Condition Evaluation

To evaluate the effectiveness in handling the explicit condition, we count the number of frames when the motion finishes under different speed settings. Fig.6 shows the result on CATER-GEN-v1 with predefined sampling interval $(3, 6)$. It can be found that the speed of generated video is consistent with ground truth. The borderline speed may cause confusion due to the sampling accuracy, but the error is within one frame. It proves that the MA is able to integrate the information specified in the explicit condition and pass it to visual tokens. We believe that this structure also works for other quantifiable conditions.

4.3.3 Composability Evaluation

In order to show the controllability of text, we try to use different descriptions to manipulate the same image. We show the results of CATER-GEN-v2 in Fig.7 trying to manipulate different objects moving to different positions. The results indicate that our model can recognize different objects as well as their attributes like "medium purple metal cone" and "large purple rubber cone". In our experiments, we observe in some rare cases that the color of objects may change in the generated videos (the cyan sphere in the last row of Fig.7). This is actually caused by the reconstruction quality of VQ-VAE. We expect that this error will disappear when a higher-capacity VQ-VAE is trained.

4.4. Diverse Video Generation

The uncertainty in description is a great challenge for TI2V task. In this experiment, we try to generate video with ambiguous descriptions to show that whether the model can generate diverse videos under the premise of semantically

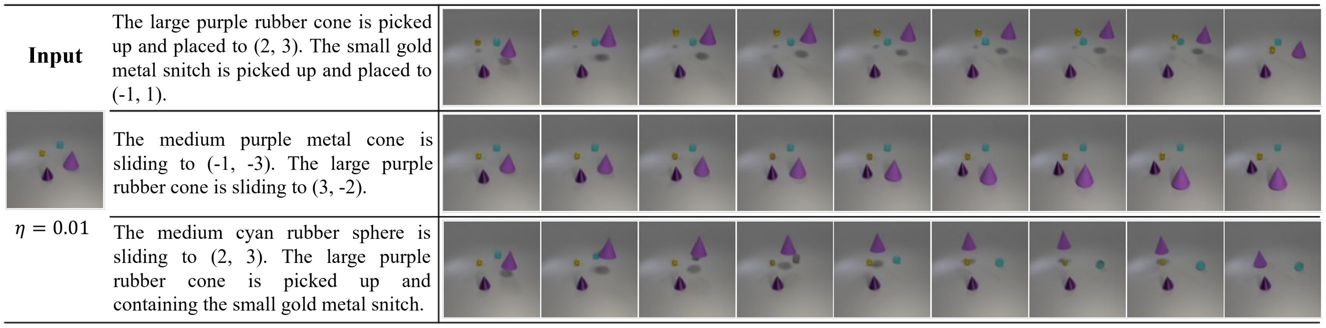


Figure 7. Generated Samples from CATER-GEN-v2 for composability evaluation. The left column shows the input with the same image and speed but different texts.

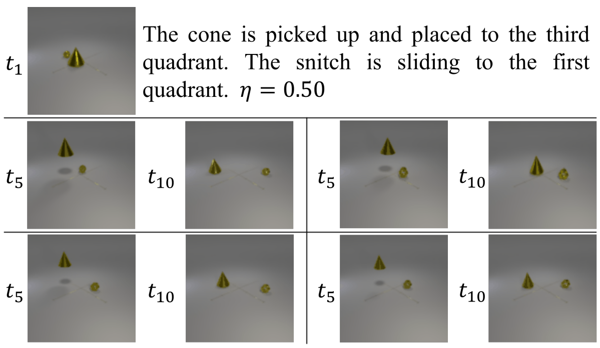


Figure 8. Generated Samples of diverse video generation from CATER-GEN-v1. Given the same image, description and speed as input in the first row, we generate four videos and show the $5th$ and $10th$ frames.

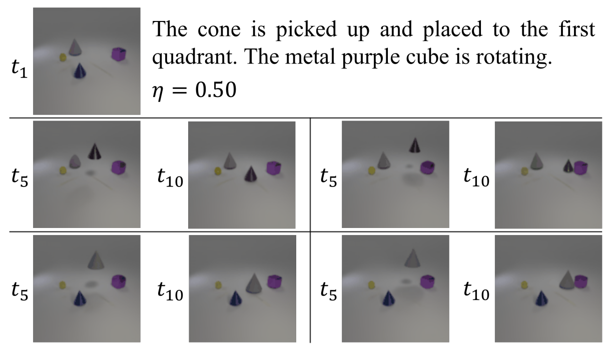


Figure 9. Generated Samples of diverse video generation from CATER-GEN-v2.

consistency with text. We conduct experiments on CATER-GEN-v1 and CATER-GEN-v2 with ambiguous descriptions and add the implicit randomness module. In the inference phrase, multiple videos are generated for the same set of image, text, and speed as input. For CATER-GEN-v1 dataset, the randomness implied in the description is the final position for action "pick-place" and "slide", as only quadrant is provided instead of exact coordinate. The results shown in Fig.8 indicate that our generation model can generate diverse videos in which objects are placed to the correct quadrant but random positions.

For CATER-GEN-v2 dataset, the implicit randomness exists in both position and referring expression. We show an example in Fig.9 that the given image contains two cones that are not distinguished in the description. The generated results show that both the gray cone and the blue cone have chances to be picked. In addition, the placement positions are all in the first quadrant but can be different coordinates.

The diverse video generation results indicate that the uncertainty existed in data distribution can be automatically extracted and integrated into motion anchor, while generating reasonable and diverse videos. It is encouraging to see the success of MAGE on these two datasets with controlled randomness. It suggests that MAGE may also work well for other challenging datasets if the randomness can be efficiently modelled. Please refer to the supplementary Sec.3 for more visualizations.

5. Conclusion

In this paper, we have introduced a novel Text-Image-to-Video generation task aiming to generate video from a static image and a text description. We have proposed MAGE which based on motion anchor enabling alignment between appearance and motion representation to generate controllable and diverse videos. To evaluate the feasibility of TI2V task and our model, we have also introduced the Modified Double Moving MNIST and CATER-GEN datasets. Experiments have shown that our model can not only generate coherent and consistent videos, but also successfully model both explicit condition and implicit randomness. However, it is still a great challenge to generate realistic and open-domain videos because of the high randomness and diversity in video-text pairs. We believe TI2V is a challenging and valuable research direction. In the future, we will extend MAGE to more challenging realistic video data.

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 4

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *ICLR*, 2018. 2

[3] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2, 2019. 3

[4] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Understanding object dynamics for interactive image-to-video synthesis. In *CVPR*, pages 5171–5181, June 2021. 1, 3

[5] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *ICCV*, pages 7608–7617, October 2019. 2

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *ICCV*, pages 5933–5942, October 2019. 1, 2

[7] Kangle Deng, Tianyi Fei, Xin Huang, and Yuxin Peng. Ircgan: Introspective recurrent convolutional gan for text-to-video generation. In *IJCAI*, pages 2216–2222, 2019. 3, 5

[8] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *CVPR*, pages 3742–3753, June 2021. 2, 3

[9] Oran Gafni, Lior Wolf, and Yaniv Taigman. Vid2game: Controllable characters extracted from real-world videos. In *ICLR*, 2020. 2

[10] Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions & temporal reasoning. In *ICLR*, 2020. 2, 5

[11] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *CVPR*, pages 7854–7863, June 2018. 1, 2

[12] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019. 5

[13] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 5

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 2

[15] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Flow-grounded spatial-temporal video prediction from still images. In *ECCV*, pages 600–615, September 2018. 2

[16] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. *AAAI*, 32(1), April 2018. 3

[17] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P. Xing. Dual motion gan for future-flow embedded video prediction. In *ICCV*, pages 1744–1752, October 2017. 2

[18] Tanya Marwah, Gaurav Mittal, and Vineeth N. Balasubramanian. Attentive semantic video generation using captions. In *ICCV*, pages 1426–1434, October 2017. 3

[19] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *CVPR*, pages 10061–10070, June 2021. 1, 2

[20] Gaurav Mittal, Tanya Marwah, and Vineeth N. Balasubramanian. Sync-draw: Automatic video generation using deep recurrent attentive architectures. In *ACM MM*, page 1096–1104, 2017. 3, 5

[21] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *CVPR*, pages 3733–3742, June 2019. 3

[22] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *ACM MM*, page 1789–1798, 2017. 3

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021. 6

[24] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, pages 2830–2839, Oct 2017. 1, 2

[25] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36, 2004. 2

[26] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, June 2018. 1, 2

[27] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, page 6309–6318, 2017. 3

[28] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, volume 29, pages 613–621, 2016. 1, 2

[29] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, pages 3332–3341, October 2017. 2

[30] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. Few-shot video-to-video synthesis. *NeurIPS*, 32:5013–5024, 2019. 1, 2

[31] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, page 1152–1164, 2018. 2

[32] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *CVPR*, pages 5264–5273, June 2020. 2

[33] Yaohui Wang, Francois Bremond, and Antitza Dantcheva. Inmodegan: Interpretable motion decomposition generative adversarial network for video generation. *arXiv preprint arXiv:2101.03049*, 2021. 2

[34] Nevan wichers, Ruben Villegas, Dumitru Erhan, and Honglak Lee. Hierarchical long-term video prediction without supervision. In *ICML*, volume 80, pages 6038–6046, July 2018. 2

[35] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 1, 3, 5

[36] Yue Wu, Rongrong Gao, Jaesik Park, and Qifeng Chen. Future video synthesis with object motion prediction. In *CVPR*, pages 5539–5548, June 2020. 2

[37] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *CVPR*, pages 2364–2373, June 2018. 2

[38] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, June 2016. 2

[39] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *ECCV*, pages 201–216, 2018. 2

[40] Jiangning Zhang, Chao Xu, Liang Liu, Mengmeng Wang, Xia Wu, Yong Liu, and Yunliang Jiang. Dtvnet: Dynamic time-lapse video generation via single still image. In *ECCV*, pages 300–315, 2020. 2, 3

[41] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris Metaxas. Learning to forecast and refine residual motion for image-to-video generation. In *ECCV*, pages 387–403, September 2018. 2