Supplementary Material
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

1. Dataset Generation

1.1. Modified Double Moving MNIST

This dataset contains 8 motion patterns combined by 4 directions and 2 modes (no bounce or bounce once). In order to improve the controllability of text and make the dataset more complex, we randomly add one static digit as background. To avoid the ambiguity of description, two digits of the same number are not allowed to exist in one image.

Following the Single/Double Moving MNIST, the combinations of digit and motion pattern in training and testing set are mutually exclusive. That means, for example, digit 9 only moves horizontally in training set, but vertically in testing test.

1.1.1 CATER-GEN-v1

CATER-GEN-v1 is a smaller and simpler version of CATER-GENs to facilitate the observation of not only actions of generated video, but also the variation of surface lighting, the shadow, and the background. There are three illuminations out of scene including key light, fill light, and back light. Once the object moves, the surface light and shallow will also change, bringing a challenge for reasonable and accurate video generation. This dataset only contains two objects: cone and snitch (like three intertwined tori) in metallic gold color. The initial position of objects on the table plane is randomly selected from a 6×6 portion. We inherit four actions in CATER: "rotate", "pick-place", "slide", and "contain". Each video randomly contains one or two actions that happen at the same time. The "rotate" action is afforded by snitch. For "pick-place" and "slide" actions, the target position is also randomly selected. We define the descriptions according to shapes, actions, and coordinates ("the cone is picked up and containing the snitch", "the snitch is sliding to (-1, 3)"). We also provide the version of ambiguous descriptions by replacing the coordinate with quadrant for diverse video generation ("the snitch is sliding to the second quadrant").

1.1.2 CATER-GEN-v2

Based on the pipeline of CATER dataset, we inherit the objects and actions. Specifically, objects include five shapes (cube, sphere, cylinder, cone, snitch), in three sizes (small, medium, large), two materials (metal, rubber) and nine colors (red, blue, green, yellow, gray, brown, purple, cyan, and the gold only for snitch). The snitch is a special object with fix size, material, and color. The "rotate" action is afforded by cubes, cylinders and the snitch, while the "contain" action is only afforded by the cones. In CATER-GEN-v2, each video randomly contains one or two actions that both start at the first frame. We fix the camera position to ensure the consistency of coordinate system. To generate explicit descriptions, we provide all properties, action, and coordinate for each moving object like "the medium blue rubber cone is picked up and placed to (1, -3)". To generate ambiguous descriptions, we randomly discard attributes (size, material, color) for each object thus brings into the uncertainty of referring expression. Like CATER-GEN-v1, we also replace the coordinate with quadrant to produce the uncertainty of movements ("the rubber cone is picked up and placed to the fourth quadrant").

2. Quantitative Results

Since video generation is a challenging and relatively new task, there are few effective metrics to evaluate generated videos currently. To quantitatively evaluate MAGE, we apply conventional pixel-based similarity metrics SSIM and PSNR for deterministic video generation. We also report several perceptual similarity metrics including image-level Fréchet Inception Distance (FID) [3] and Learned Perceptual Image Patch Similarity (LPIPS) [2], as well as video-level Fréchet-Video-Distance (FVD) [9]. To evaluate the diversity of generated videos given ambiguous text, following previous work [1], we measure the average mutual distance of generated video sequences in the feature space of

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

Mode	Datasets	FID ↓	LPIPS ↓	FVD ↓	DIV VGG ↑	DIV I3D ↑
Deterministic (explicit text)	CATER-GEN-v1	62.66	0.20	31.70	0	0
	CATER-GEN-v2	39.56	0.20	57.55	0	0
Diverse (ambiguous text)	CATER-GEN-v1	62.89	0.22	45.49	0.15	0.45
	CATER-GEN-v2	39.38	0.26	69.44	0.37	2.06

Table 1. Qualitative results on CATER-based datasets under deterministic and diverse video generation, respectively.

both VGG-16 [7] and I3D [8] network. The VGG and I3D backbones used in similarity and diversity metrics are pre-trained on ImageNet [5] and Kinetics [4], respectively.

Datasets	SSIM ↑	PSNR ↑	
		VQ-VAE	MAGE
Single Moving MNIST	0.97	43.12	33.89
Double Moving MNIST	0.87	38.80	24.66
Modified Moving MNIST	0.85	37.63	23.24
CATER-GEN-v1	0.97	47.01	35.03
CATER-GEN-v2	0.95	40.21	32.74

Table 2. Qualitative results under deterministic video generation.

Due to the novel setting of TI2V task, there exists a restrictive relation between high similarity and large diversity. Generated videos are required to be more diverse and semantically consistent with text at the same time. It is hard to fairly compare with other methods for I2V or T2V task, as methods for I2V task fail to generate controllable video with complicated motion in text. And methods for T2V task tend to generate correlated visual features that have been seen in the training stage, making it difficult to generate video from a novel image and model the uncertainty in text.

Tab.2 shows the PSNR and SSIM results on all datasets under deterministic video generation that only involves explicit text descriptions. In the testing stage, the speed η is randomly sampled from $(0, 1)$ for each sample. As the video generation performance is based on the reconstruction accuracy of VQ-VAE, both PSNR results of reconstructed video form VQ-VAE only and generated videos from MAGE are reported against ground truth videos. It can be found that our generated videos achieve high similarity with ground truth videos. When the dataset becomes harder, the video generation performance is considerably affected by VQ-VAE and declines.

To further evaluate the ability to handle ambiguous text, we compare the perceptual similarity and diversity after applying implicit randomness module. When calculating diversity metric, we fix the speed input and generate 5 video sequences for each sample. The results are shown in Tab.1. Given explicit text, the generated video is unique and shows high similarity with reference video. After replacing explicit text with ambiguous text, the ground-truth video is no longer unique in this situation. The similarity between generated video and reference video decreases within an acceptable range. At the same time, the model is able to generate diverse videos. The results of CATER-GEN-v2 show much higher diversity than CATER-GEN-v1, which is also consistent with different degrees of uncertainty in text.

In addition, to validate the effectiveness of the cross-attention of MA and axial transformers, we conduct ablation study to compare with concatenation operation and vanilla transformer, respectively, in Tab. 3. First, the 1st and the 2nd rows show that cross-attention in MA has better performance than concatenation with 1.4 decrease on FID while keeping similar LPIPS and FVD. Then, comparing the 2nd and the 3rd rows shows that axial transformers incur some performance loss but it remarkably reduces the computational complexity of vanilla transformers by 46%, which is consistent with our expectation.

Transformers		MA		FID	LPIPS	FVD	GFLOPs↓	Throughput↑
Vanilla	Axial	Concat	Cross	↓	↓	↓		
√		√		63.4	0.20	30.4	85.1	12.3
√			√	62.0	0.20	30.5	96.8	11.5
	√		√	62.7	0.20	31.7	52.7	37.2

Table 3. Ablation study under deterministic video generation on CATER-GEN-v1.

3. Additional Visualizations

3.1. Attention Visualization in Motion Anchor

To show whether the motion anchor locates right objects and their motion, we visualize the attention map in cross-attention when generating motion anchor. Since the semantics of motion and object information in text have interacted and fused in the front text encoder, we select integral noun phrase composed of 4 attributes to show the response in image. As shown in Fig.2, with visual token embeddings as query, we average the attention maps of different heads and show the mean attention weights of specified noun phrase (marked with same color in text). The visualizations show that the cross-attention operation is aware of multiple ob-

jects in the scene and locates the specified objects.

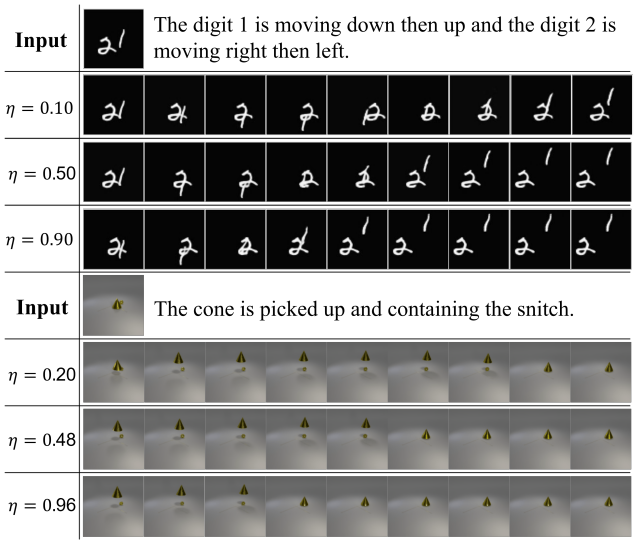

Figure 1. Generated Samples from Modified Double Moving MNIST and CATER-GEN-v1 for explicit condition evaluation. The input row is the given image and description. The left column is the input speed.

3.2. Visualization of Explicit Condition

To visualize the effect of explicit condition speed, we give the same image and description but input different speeds. Examples are shown in Fig.1. Suppose each video contains 20 frames and the predefined sampling interval in training is $(1, 2)$, then $\eta = 0.50$, for example, stands for corresponding sampling interval with 1.5. By giving different η, it can be found that the model correctly generates videos with corresponding speed. More generated videos from CATER-GEN-v2 are shown in Fig.3.

3.3. Visualization of Composability

More compositional video generation results from Modified Double Moving MNIST and CATER-GEN-v1 are shown in Fig.4. Given an image and a fixed speed, three descriptions are input separately to specify different objects and actions. Results show correct concordance with text both on moving targets and actions.

3.4. Visualization of Implicit Randomness

We also show diverse generated videos in Fig.5 with ambiguous text as input. The $5th$ frames and $10th$ frames from two generated video with the same input are shown. It can be found that, even given a difficult image and a complicated caption, our method can model the implied randomness (including final position or action subject) and generate diverse and relatively satisfactory results. However, when the appearance of image is too complicated, there may be deformation and clipping problems (like the last row). Besides, since the VQ-VAE is trained on frame-level, there are some objects in reconstructed videos have color change. That will result in color inconsistency of generated videos (like the third row on the right).

3.5. Failure Cases

We also show some failure examples in Fig.6. The first example shows the effect of VQ-VAE. In our method, VQ-VAE is responsible for image tokenizer before generation (from 128×128 to 16×16) and reconstruction after generation. Therefore, the performance of VQ-VAE will directly affect video generator. Because of the large down-sampling ratio, VQ-VAE may lose fine-grained texture and result in wrong shape of reconstructed object (the gray sphere in the $6th$ frame of reconstruction-only video), further leading to distortion of generated video (the gray sphere in the $4th$ frame of generated video).

On the other hand, since MEGA generates video tokens at latent space with spatial size 16×16, the small resolution makes it hard to split two objects with overlap (like the blue sphere and yellow cube in the second example). This may also cause distortion when two objects intersect during moving.

3.6. Visualization of Realistic Video Generation

Except for synthetic videos, we also wonder whether our model can generate realistic videos. However, existing paired video-text datasets contain high uncertainty and much noise and like scene change or the emergence of irrelevant objects. The texts are also not fine-grained enough, making video generation hard to control. Therefore, we evaluate our method on KTH [6] which is a relatively clean action recognition dataset. The KTH dataset contains 2391 video clips of six human actions performed by 25 people in four different scenarios. We use the original train-test split. The text is formed like "A person is $[action\_label]$." The reference video, reconstructed video of VQ-VAE only and generated video of MAGE are shown in Fig.7. Compared to reference videos, the reconstructed videos are blurrier and lost details due to the limitation of VQ-VAE. Even though, generated videos do not have much degradation in quality compared to reconstructed videos. The speed of generated videos is also consistent with input. However, for action like "walking", there may be no person in the first given image, resulting in that the model tends to learn an "average person" (like a body with black color) in the training stage. This may cause the body to gradually turn black during generation. It also shows that generating unseen objects is still a challenge.

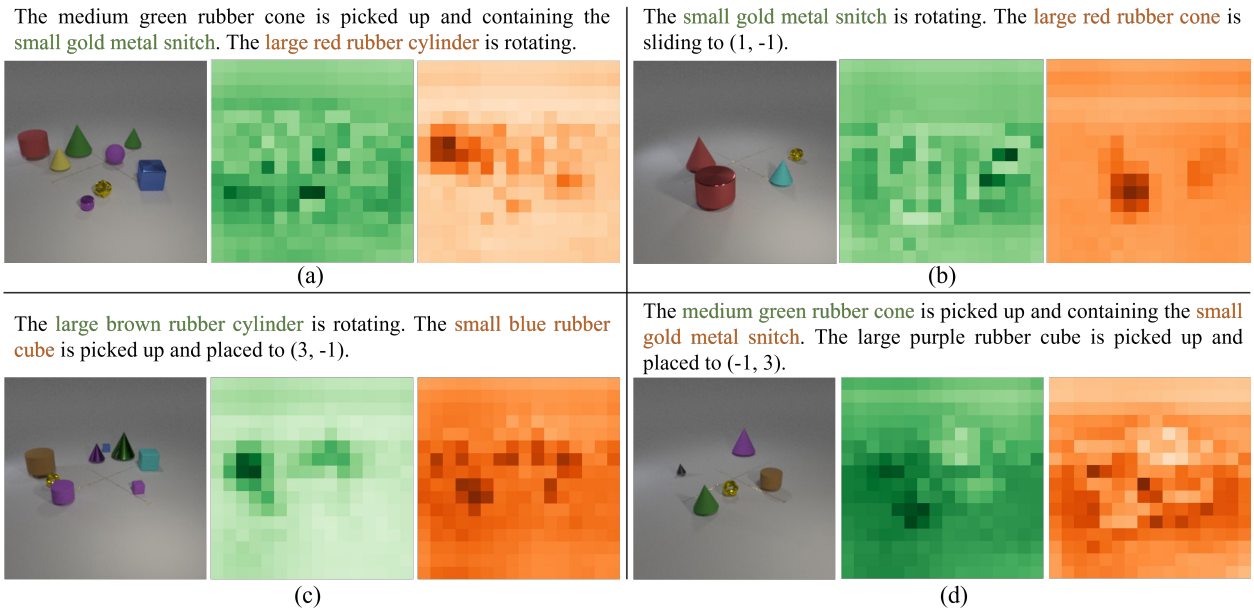

Figure 2. Visualization of attention weights from cross attention during generating motion anchor. With a group of visual token embeddings (16×16) from the corresponding position in image as query, each element in attention map stands for the mean weight of 4 word embeddings of specified noun phrase marked with same color in text. The darker the color, the greater the response.

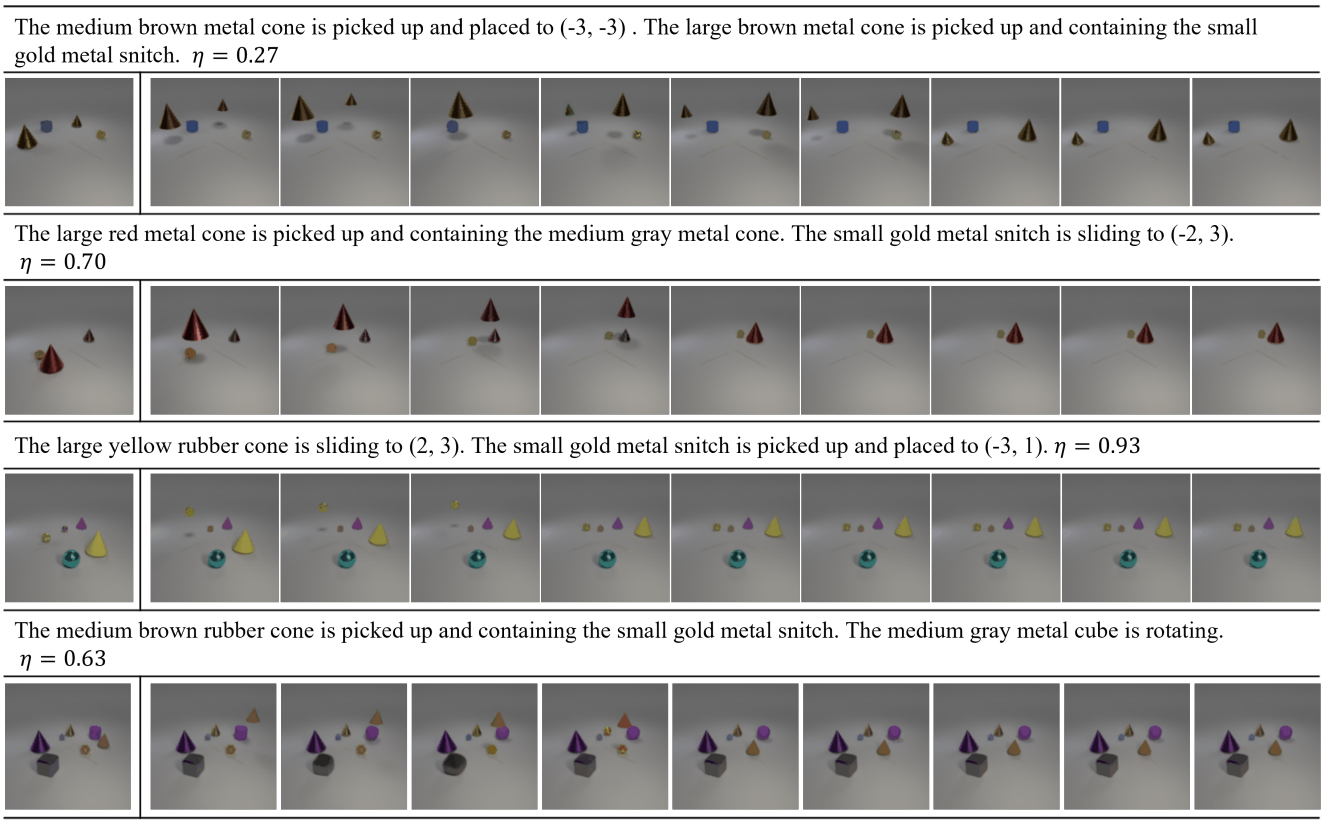

Figure 3. Generated Samples from CATER-GEN-v2 under deterministic video generation.

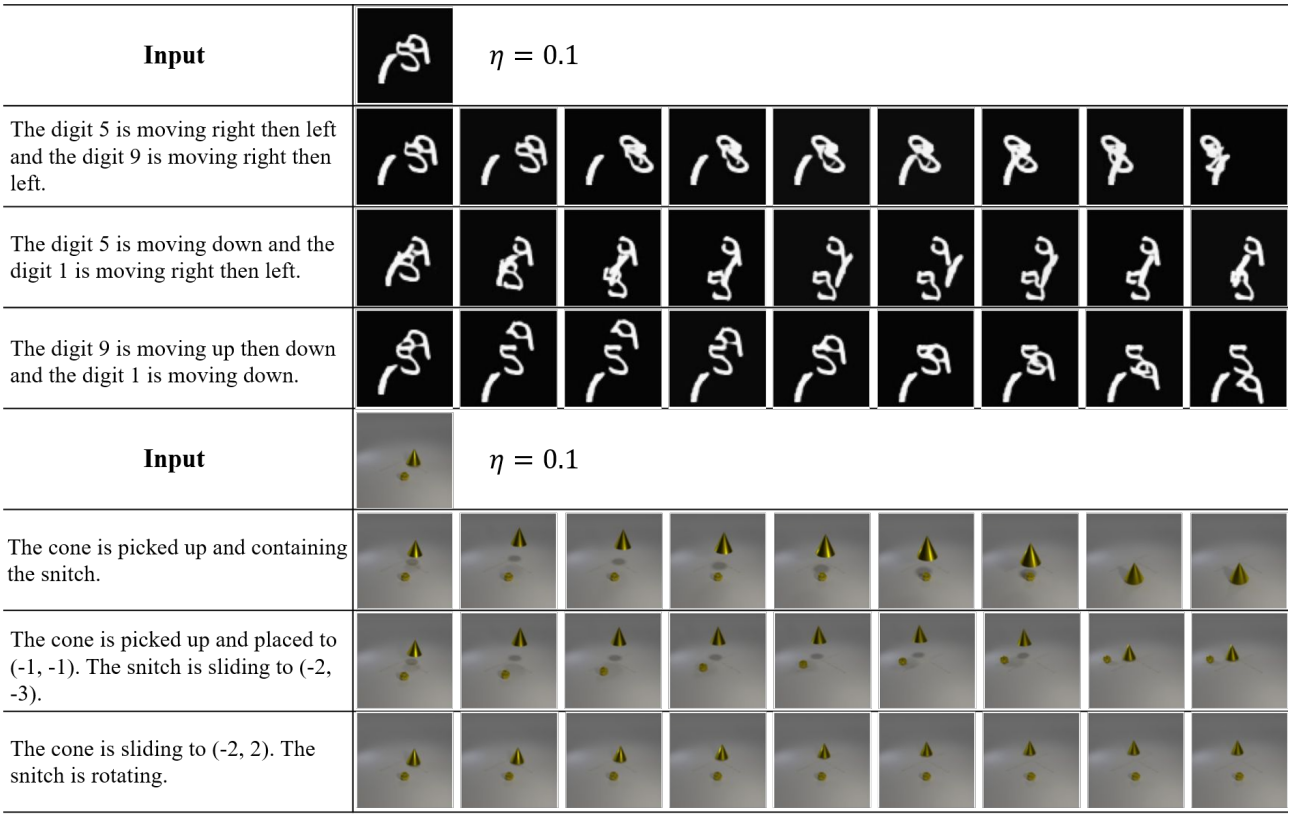

Figure 4. Generated Samples from Modified Double Moving MNIST and CATER-GEN-v1 for composability evaluation.

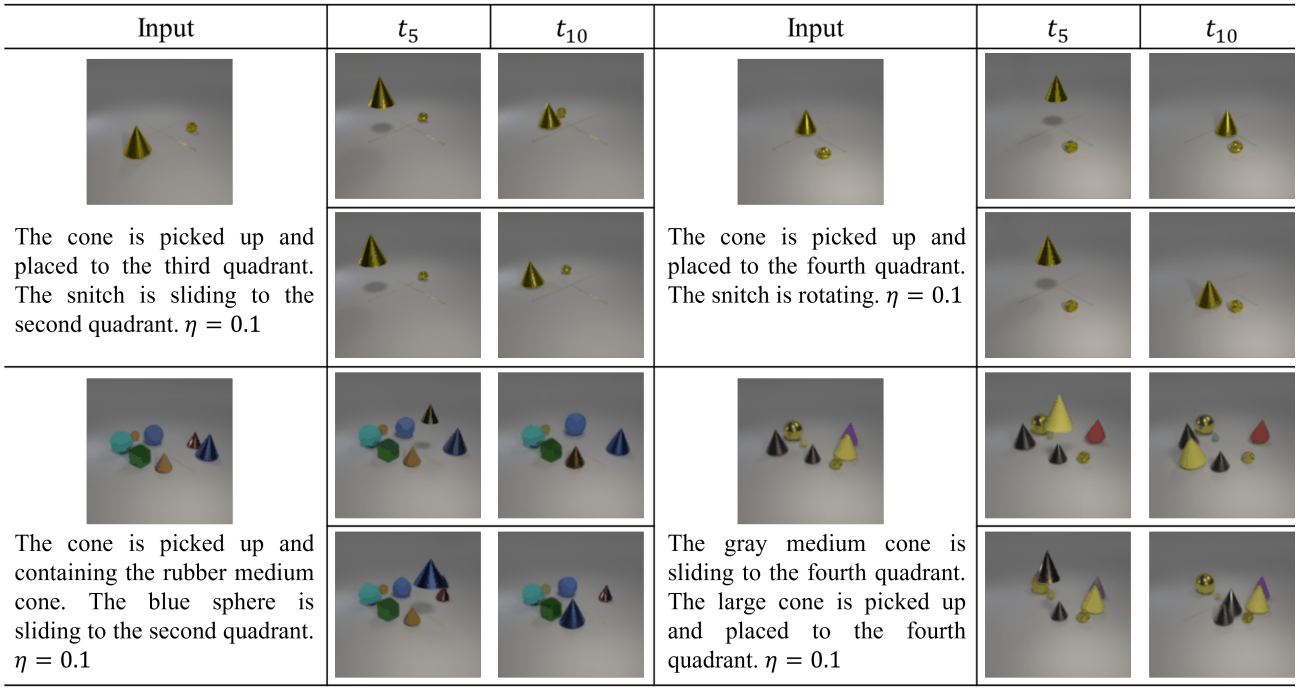

Figure 5. Generated Samples from CATER-GEN-v1 and CATER-GEN-v2 for diverse video generation.

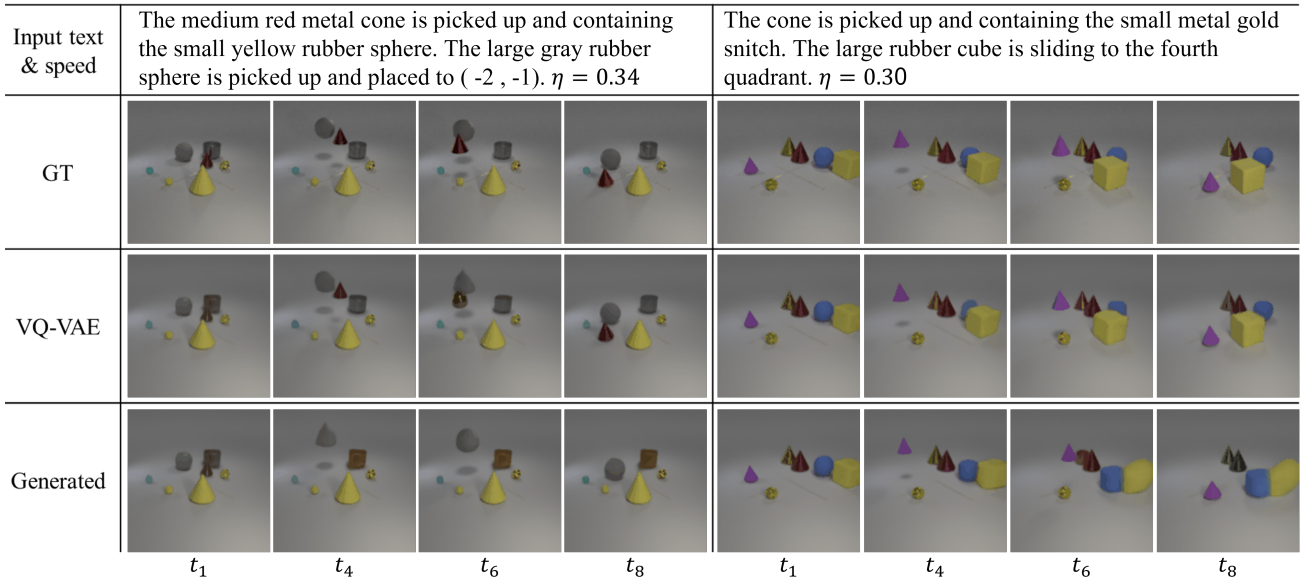

Figure 6. Failure Cases from CATER-GEN-v2. Four rows from top to bottom are input text and speed, the reference video, the reconstructed video from VQ-VAE only, and the generated video from MEGA, respectively. For each video, the $1st$, $4th$, $6th$, and $8th$ frames are shown.

4. Limitations

Although our method can generate controllable and diverse videos, there are failure cases that reflect some limitations of MAGE. Besides, TI2V task also faces demands in evaluation metrics and datasets. We summarize those limitations here.

- **VQ-VAE based Architecture:** Despite VQ-VAE greatly reduces the data volume and facilitates the training of video generator, the reconstruction performance of VQ-VAE will directly affect prediction accuracy of subsequent video generator. Especially under a large down-sampling ratio, VQ-VAE may lose fine-grained texture, leading to distortion of object. Besides, as VQ-VAE is trained on image-level, the temporal consistency is not guaranteed, which may result in inconsistency of predicted videos. Therefore, higher performance VQ-VAE and proper down-sampling ratio should be considered to help model generate more consistent and high-resolution videos.

- **Evaluation Metrics:** Evaluating the quality of generated videos is challenging for TI2V task, especially for ambiguous text descriptions. Since the "correct" video may not be unique and accessible, existing similarity metrics (e.g. PSNR, LPIPS, FVD) measuring the semantic consistency between generated video with one of "correct" videos are not accurate. Meanwhile, diversity metrics (e.g. DIV VGG/I3D) can only reflect the variation between generated videos no matter whether they are consistent with text descriptions. Thus, appropriate metrics to evaluate both accuracy and diversity are needed.

- **Realistic Video Generation:** The difficulties for realistic video generation lie in not only the great uncertainty in realistic videos, but also the lack of appropriate datasets. Most of existing video-text paired datasets are composed of coarse-grained text descriptions, making it harder to generate coherent motion given the first image. Generating realistic and open-domain videos is still a major challenge of TI2V task.

References

[1] 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. 1

[2] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *NeurIPS*, 29:658–666, 2016. 1

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 1

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,

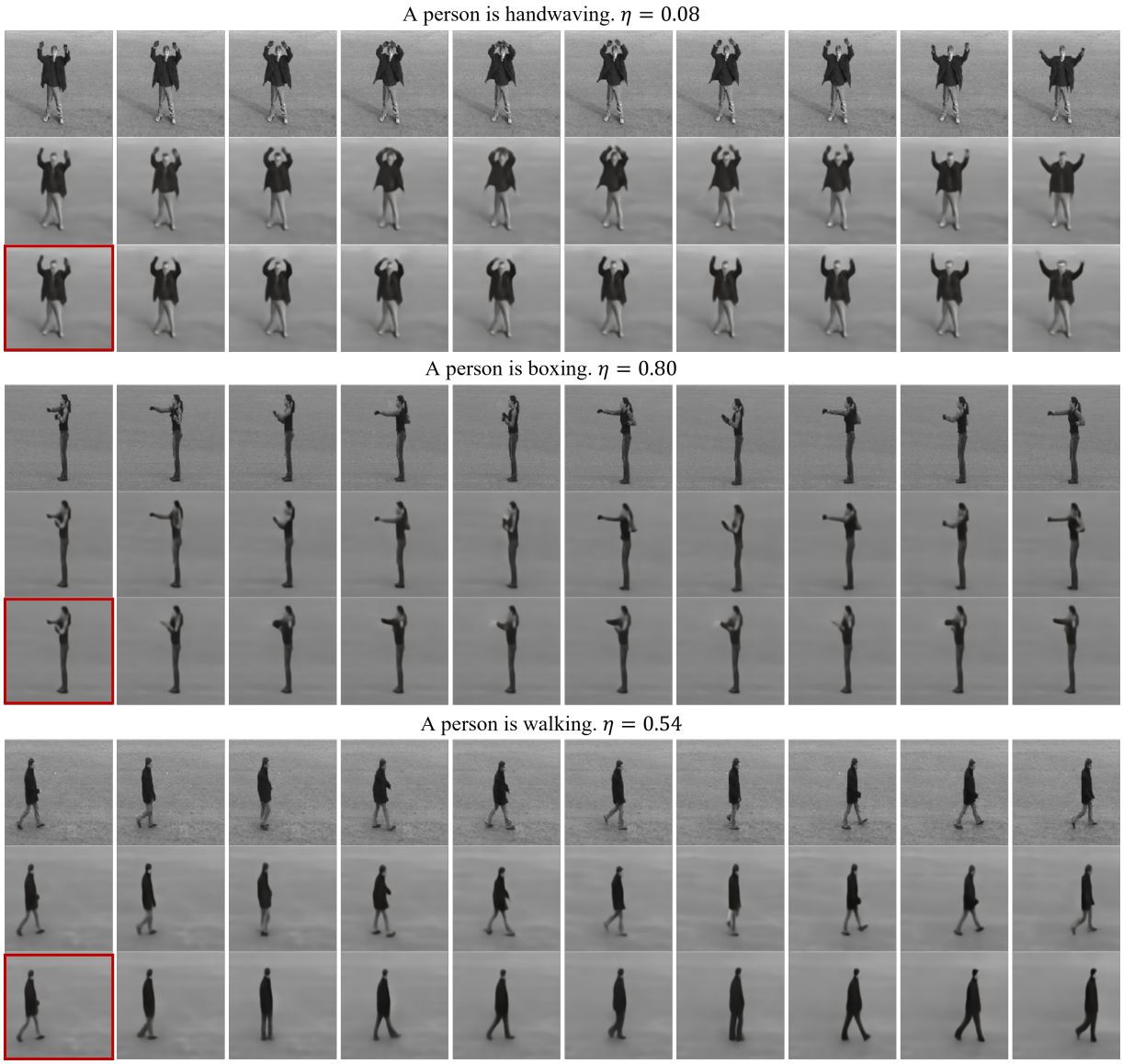

Figure 7. Generated Samples from KTH. Three rows from top to bottom of each example are the reference video, the reconstructed video from VQ-VAE only, and the generated video from MEGA, respectively. The red box represents the first given image when generating video.

Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 2

[6] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, volume 3, pages 32–36, 2004. 3

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[9] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1