

# Supplementary Materials for TI2V-Zero: Zero-Shot Image Conditioning for Text-to-Video Diffusion Models

Haomiao Ni<sup>1\*</sup>    Bernhard Egger<sup>2</sup>    Suhas Lohit<sup>3</sup>    Anoop Cherian<sup>3</sup>    Ye Wang<sup>3</sup>  
Toshiaki Koike-Akino<sup>3</sup>    Sharon X. Huang<sup>1</sup>    Tim K. Marks<sup>3</sup>

<sup>1</sup>The Pennsylvania State University, USA

<sup>2</sup>Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany

<sup>3</sup>Mitsubishi Electric Research Laboratories (MERL), USA

<sup>1</sup>{hfn5052, suh972}@psu.edu    <sup>2</sup>bernhard.egger@fau.de    <sup>3</sup>{slohit, cherian, yewang, koike, tmarks}@merl.com

<https://merl.com/demos/TI2V-Zero>

## A. Dataset Details

We conduct extensive experiments on three diverse datasets, including facial expression dataset MUG, action recognition dataset UCF101, and our self-created dataset OPEN. Here we present comprehensive details about these datasets.

For the MUG dataset, we randomly select 5 male and 5 female subjects from the available 52 individuals, and 4 expressions from the provided 7 expression classes. Detailed information about selected subjects and corresponding expression labels are presented in Tab. 1. To convert expression class labels to text prompts for input, we use the following templates: “A woman with the expression of slight {label} on her face.” for female subjects, and “A man with the expression of slight {label} on his face.” for male subjects. Considering that the average original video length on MUG is about 72 frames, we uniformly sample 16 frames from most of the videos to create the real videos. For videos with more than 80 frames, we sample the videos every 5 frames until we obtain 16 frames to form the real videos.

For the UCF101 dataset, we initially randomly select some action classes from the provided 101 classes. Subsequently, we identify and choose 10 action classes where both ModelScopeT2V and VideoCrafter1 are able to synthesize promising videos. Table 2 shows the details of selected action class labels and their corresponding text prompts. For each action class, we simply choose the first 10 subjects. Given that the average original video length on the UCF101 dataset is approximately 200 frames, we sample the videos every 10 frames until 16 frames are obtained to form the real videos. For videos containing less than 160 frames, we uniformly sample 16 frames.

For the OPEN dataset, we first employ ChatGPT 3.5<sup>1</sup>

Male ID	007, 010, 013, 014, 020
Female ID	001, 002, 006, 046, 048
Expression	Anger, Happiness, Sadness, Surprise

Table 1. Details of selected subjects and expression classes on the MUG dataset.

Action Class	Text Prompt
ApplyEyeMakeup	“A person is applying eye makeup.”
BabyCrawling	“A baby is crawling.”
BreastStroke	“A person is performing breaststroke.”
Drumming	“A person is drumming.”
HorseRiding	“A person is riding horse.”
Kayaking	“A person is kayaking.”
MilitaryParade	“Military parade.”
PlayingGuitar	“A person is playing guitar.”
Surfing	“A person is surfing.”
ShavingBeard	“A person is shaving beard.”

Table 2. Details of selected action class labels and corresponding text prompts on the UCF101 dataset.

to generate 10 text prompts by inputting the query “Could you randomly generate 10 text prompts for testing text-to-video models?”. We list these 10 text prompts in Tab. 3. Then we use Stable Diffusion 1.5 with the model ID `dreamlike-photoreal-2.0`<sup>2</sup> to generate 100 images for each of 10 text prompts, resulting in a total of 1,000 images as starting frames.

## B. Additional Experimental Results

**More Prior Work Comparisons.** In Tab. 4, we conduct additional experiments to compare our proposed model with the open-domain TI2V model VideoComposer [1] on MUG and UCF101 datasets, where our TI2V-Zero achieves supe-

<sup>\*</sup>Work done during an internship at MERL.

<sup>1</sup><https://openai.com/blog/chatgpt>

<sup>2</sup><https://huggingface.co/dreamlike-art/dreamlike-photoreal-2.0>

1	“A mesmerizing display of the northern lights in the Arctic.”
2	“A bustling street market in Marrakech with colorful textiles and spices.”
3	“A futuristic cityscape with holographic advertisements and flying cars.”
4	“A romantic gondola ride through the canals of Venice at sunset.”
5	“A group of friends on a road trip, singing along to their favorite songs.”
6	“A serene mountain cabin covered in a fresh blanket of snow.”
7	“A thrilling skateboarder performing tricks in a skate park.”
8	“A bustling night market in Bangkok with street food vendors and live music.”
9	“A high-speed bullet train racing through a scenic countryside.”
10	“A group of explorers uncovering the mysteries of an ancient temple in the jungle.”

Table 3. The 10 text prompts used in the OPEN dataset.



Figure 1. Examples of generated video frames in video prediction task conditioning on different numbers of given images. The 1st, 6th, 11th, and 16th frames of each output video are shown in each column. Each generated video has 16 frames with a resolution of  $256 \times 256$ . 1 image, 4 images, 8 images indicate the use of the first 1, 4, and 8 real video frames in the ground truth video to predict the next 15, 12, and 8 frames, respectively.

rior performance.

**Extension to Video Prediction Task.** We have presented the results of video infilling and long video generation in the main paper. In Fig. 1 and our supplementary videos, we show the application of our proposed TI2V-Zero to the video prediction task. Specifically, we conduct experiments using the first 1, 4, and 8 real video frames from the ground truth videos to generate 16-frame videos, i.e., synthesize the subsequent 15, 12, and 8 frames, respectively. As illustrated in Fig. 1, when only 1 image is provided, the woman in the generated video applies the powder brush to the eye differently from the real video. With 4 images, the woman in the synthesized video applies the brush to the

Model	MUG			UCF101	
	FVD↓	sFVD↓	tFVD↓	FVD↓	tFVD↓
VideoComposer [1]	1899.08	2294.60±482.45	2050.69±116.92	633.32	1606.13±355.87
TI2V-Zero (Ours)	<b>180.09</b>	<b>267.17±74.72</b>	<b>252.77±39.02</b>	<b>477.19</b>	<b>1306.75±271.82</b>

Table 4. Quantitative comparison between VideoComposer and TI2V-Zero (w/ Resample) for TI2V generation.

DDIM	Resample	Time (s)
10	0	5.46
50	0	24.86
10	2	14.90
10	4	24.70

Table 5. The average inference time for generating one frame using our proposed TI2V-Zero under different sampling settings. The terms DDIM and Resample represent the number of steps of using DDIM sampling and resampling.

same eye as in the ground truth video, but it is still hard to maintain the same appearance of the brush as real video. When extending to 8 frames, the model can synthesize a video that is consistent with the given real video.

**Inference Time and GPU Usage.** In Tab. 5, we report the average inference time of generating one frame with our proposed TI2V-Zero under different sampling settings, when using a batch size of one on a Quadro RTX 6000 GPU. The GPU usage for each setting is 9,885 MB. With the same GPU, the baseline DynamiCrafter takes about 155 seconds to generate a 16-frame video using their default settings.

## C. Discussion with Concurrent Work

A concurrent work to ours, AnimateZero [2], also adopts a similar repeating operation. However, we are different in several aspects. In our framework, when computing temporal attention outputs, the sources of keys are derived either from the given image or previously synthesized images, whereas AnimateZero relies on keys from the given image or noise. Moreover, AnimateZero shares keys and values from spatial self-attention of the first frame across the other frames; this may make it hard to generate large motions and novel scenes, as the content is constrained to the informa-

tion available in the first frame. In contrast, our framework demonstrates the ability to generate promising videos containing intricate motions with input images of various styles across a wide variety of scenes.

## D. Information about Example Videos

We include eight MP4 files of example video clips generated by our proposed method in the Supplementary materials.

- **mug.mp4** includes the video clips generated by the state-of-the-art model DynamiCrafter and our proposed TI2V-Zero for 4 expressions of 4 subjects from the MUG dataset.
- **ucf.mp4** contains the synthesized video clips produced by DynamiCrafter and our TI2V-Zero for action classes from the UCF101 dataset.
- **open.mp4** contains the generated video clips using DynamiCrafter and our TI2V-Zero for 10 text prompts from the OPEN dataset.
- **ablation.mp4** compares the generated video clips under different sampling strategies using our proposed TI2V-Zero on the MUG dataset. The terms `Inversion`, `DDIM`, and `Resample` denote the application of DDPM-based inversion, the steps using DDIM sampling, and the iteration number using resampling, respectively.
- **long\_video.mp4** displays one example video clip showing the application of our proposed TI2V-Zero to generate a 128-frame long video.
- **prediction.mp4** shows one example video clip illustrating the application of our proposed TI2V-Zero to video prediction task, conditioning on different numbers of given images.
- **motivation.mp4** shows the video clips generated by the replacing-based baseline approach and our proposed TI2V-Zero for different video tasks (corresponding to Fig. 3 in our main paper).
- **intricate.mp4** shows the video clips generated with intricate text and image inputs, including two (16-frame) videos and one long (64-frame) video. Each first-frame image in the video clips was generated by Stable Diffusion 1.5.

## References

- [1] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 1, 2
- [2] Jiwen Yu, Xiaodong Cun, Chenyang Qi, Yong Zhang, Xintao Wang, Ying Shan, and Jian Zhang. Animatezero: Video diffusion models are zero-shot image animators. *arXiv preprint arXiv:2312.03793*, 2023. 2