Tell Me What Happened: Unifying Text-guided Video Completion via Multimodal Masked Video Generation

Tsu-Jui Fu[1], Licheng Yu[2], Ning Zhang[2], Cheng-Yang Fu[2],
Jong-Chyi Su[3], William Yang Wang[1], Sean Bell[2]
[1]UC Santa Barbara [2]Meta [3]NEC Laboratories America
{tsu-juifu, william}@cs.ucsb.edu jcsu@nec-labs.com
{lichengyu, ningzhang, chengyangfu, seanbell}@meta.com

Abstract

Generating a video given the first several static frames is challenging as it anticipates reasonable future frames with temporal coherence. Besides video prediction, the ability to rewind from the last frame or infilling between the head and tail is also crucial, but they have rarely been explored for video completion. Since there could be different outcomes from the hints of just a few frames, a system that can follow natural language to perform video completion may significantly improve controllability. Inspired by this, we introduce a novel task, text-guided video completion (TVC), which requests the model to generate a video from partial frames guided by an instruction. We then propose Multimodal Masked Video Generation (MMVG) to address this TVC task. During training, MMVG discretizes the video frames into visual tokens and masks most of them to perform video completion from any time point. At inference time, a single MMVG model can address all 3 cases of TVC, including video prediction, rewind, and infilling, by applying corresponding masking conditions. We evaluate MMVG in various video scenarios, including egocentric, animation, and gaming. Extensive experimental results indicate that MMVG is effective in generating high-quality visual appearances with text guidance for TVC.

1. Introduction

Generative video modeling [15, 70, 84] has made great progress, which first succeeds in unconditional video generation [40, 64]. More recently, video prediction [28, 36, 47] has been trying the controllable setting, which anticipates the future by completing a video from the past frames or a static starting image [37, 97]. However, video prediction may produce various outcomes, which makes it difficult to meet human expectations. For the example in Fig. 1(a), the game agent can keep jumping to the right or move back and turn left. The limited guidance from only the first frame is insufficient to tell the intention. For humans, language is the most straightforward way of communication. If a system can follow an instruction to accomplish video completion, it will significantly improve its controllability and make a vast application impact. On the other hand, compared with video prediction, video rewind and infilling have been rarely studied [39, 79], but they are also crucial. Breaking the limitation of chronological guidance should make the visual guidance more flexible, leading to a general video completion.

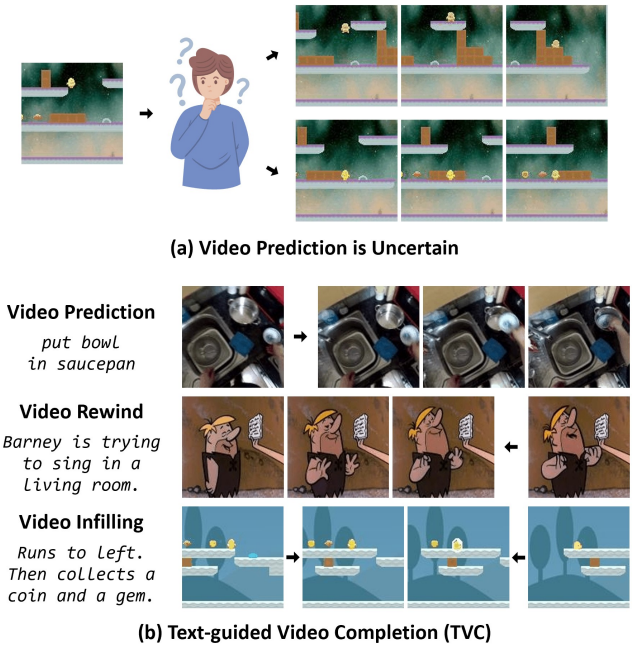

Figure 1. The introduced text-guided video completion (TVC) task. (a) Video prediction may have different outcomes without text guidance. (b) TVC performs video completion from the first frame (**prediction**), the last frame (**rewind**), or both (**infilling**), guided by the textual description.

We thus introduce a novel task, text-guided video completion (TVC), where the partial frames and a given instruction jointly guide the video generation. As illustrated in

Fig. 1(b), we consider three scenarios of video completion: **prediction** from the first frame, **rewind** from the last frame, and **infilling** between the head and tail. The missing (to-be-completed) event should follow the textual instruction. Compared to generating content from scratch [43, 87], TVC requests models to understand the given visual and textual guidance before generation, which better mimics how human imagines after seeing and listening in our daily lives.

To tackle TVC, we present Multimodal Masked Video Generation (MMVG) to perform video completion. Specifically, we represent the video frames as discrete visual tokens by temporal-aware VQGAN [54, 76]. One key challenge is to deal with the video frames that are not presented in chronological (*e.g.*, the last frame for rewind). Different from autoregressive models [23, 86] that only condition on the previous frames, MMVG carries out video completion in an encoder-decoder manner. Specifically, we propose a masking strategy that masks different parts of the video and feeds them as the input to the multimodal encoder with the instruction. As shown in Fig. 2, we allow MMVG to consider the visual hints from different time points, and the decoder learns to produce the full target video. By varying the masking conditions (including the cases of only the first or last frame being accessible), a single MMVG can address all TVC tasks, including video prediction, rewind, and infilling. Moreover, learning the recovery from partial frames also empowers MMVG with a strong temporal coherence, contributing to better generative video quality.

We consider videos in diverse scenarios for the TVC evaluation. There are Kitchen [13], Flintstones [26], and MUGEN [29] corresponding to the egocentric, animation, and gaming scenes. The model should generate videos such as performing kitchen activities in the first-person view, making characters act the assigned behavior, or imitating an agent playing game. All should be guided with the first/last (or both) frame(s) and controlled through the given human instructions. We also compare MMVG with previous methods [23, 57, 94, 96] on UCF-101 [69] and BAIR [18] for the classic video generation/prediction tasks.

Experimental results demonstrate that instruction is necessary to make video completion controllable, MMVG can address all three TVC tasks, and our proposed masking strategy enhances the temporal modeling, which further benefits general video generation/prediction. In summary, our contributions are three-fold:

- We introduce TVC to generate a video from partial frames and control the temporal dynamics via natural language, where our video completion includes 3 cases: prediction, rewind, and infilling.
- We propose MMVG with an effecitve masking strategy to address all TVC tasks through a single training.
- Extensive experiments show that our MMVG can handle various types of video completion as well as video generation/prediction. We believe TVC can become a new topic in vision-and-language research.

2. Related Work

Video Generation/Prediction. Video generation aims to synthesize diverse videos from latent inputs [1, 74, 80]. Various generative modelings have shown promising results, including generative adversarial networks (GAN) [10, 24, 72, 96], autoregressive transformers [23, 77, 94], and denoising diffusion models [16, 33, 34]. Upon that, video prediction [2, 3, 25, 57, 85], which considers past frames to anticipate future observations, should maintain temporal dynamics from static images. Though the overall idea is also to complete a video from partial frames, other tasks, such as rewind and infilling [39, 79, 90], are not extensively explored. In this paper, we introduce TVC to comprehensively investigate the ability of video completion and make it more maneuverable via textual description.

Text-to-Image/Video Genreation. Generating visual content from language [9, 50, 71] has a vast application value in creative visual design. Previous works rely on the GANs framework [49] to produce images [19, 20, 22, 55, 61, 92] or videos [43, 46, 52], conditioned on text. With large-scale datasets [4, 65, 66, 83], recent pre-trained models can generate high-quality natural images from open-domain description through discrete visual tokens [12, 17, 54, 59, 76, 95] or the diffusion process [51, 58, 62, 63]. Leveraging such techniques further extends to generate vivid videos [32, 35, 68, 78, 86, 87]. However, those methods that depend on autoregressive generation can only be guided chronologically [27, 38]. Besides, video diffusion models require a deterministic video length, which cannot consider diverse temporal durations. In contrast, MMVG can perform video completion in different lengths from arbitrary time points and address all TVC tasks just with a single training.

Text-guided Video-to-Video. Video inpainting [7, 41, 91], segmentation reconstruction [81, 82], or video style transfer [8, 14, 88] can be seen as a particular case of video-to-video synthesis (V2V). Even if text-guided V2V [5, 21, 93] can be controlled by language, it is still conditioned on a full video, where the temporal dynamics are usually provided. Different from that, TVC requires to regain the missing event from just partial guidance. It is more challenging since the model has to capture what happened from the instruction, maintain the temporal coherence among limited frames, and produce a complete video.

3. Text-guided Video Completion (TVC)

3.1. Task Definition

We study the text-guided video completion (TVC) task to perform video completion from the first frame (prediction),

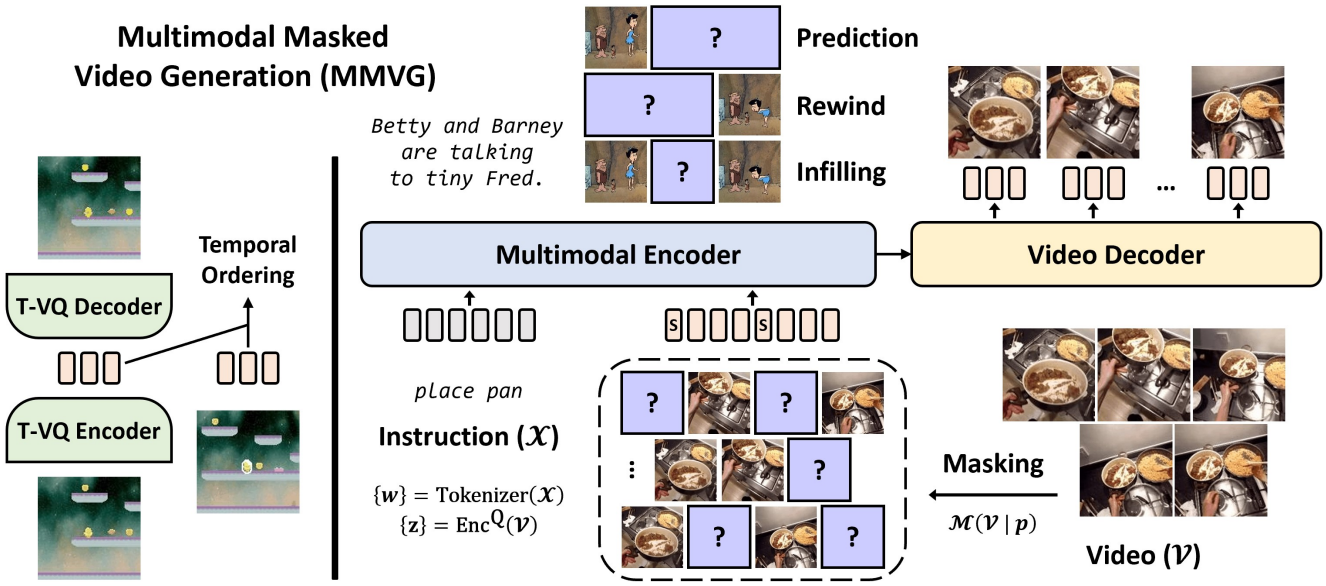

Figure 2. An overview of our Multimodal Masked Video Generation (MMVG). We present temporal-aware VQGAN (T-VQ) for discrete visual representation. MMVG considers the instruction $\mathcal{X}$ and partial frames of the video $\mathcal{V}$ from diverse time points through masking, learning to generate the complete $\mathcal{V}$. In this way, a single trained MMVG can perform all prediction, rewind, and infilling tasks.

the last frame (rewind), or the head and tail (infilling), conditioned on the textual instruction. During training, we have pairs of videos $\mathcal{V}$ and corresponding instructions $\mathcal{X}$. Specifically, $\mathcal{V}$ consists of N frames as $\{v_1, v_2, ..., v_N\}$. Our goal is to train a unified model that generates the complete $\mathcal{V}$ given the partial frames from arbitrary time points and $\mathcal{X}$.

3.2. Multimodal Masked Video Generation

Overview. An overview of our Multimodal Masked Video Generation (MMVG) is illustrated in Fig. 2. To model the video along with language, we propose temporal-aware VQGAN to represent a frame as visual tokens, which converts it into the same discrete space as the words. We present an effective masking strategy that masks different video parts for video completion learning. Those missing fragments are replaced by the unique [SPAN] tokens, and we consider the visual guidance from diverse time points. The multimodal encoder consumes the text and the partial-missing video, and the decoder learns to produce the complete video from arbitrary guided frames. By varying the masking conditions, MMVG learns to utilize the [SPAN] token and unifies all TVC tasks during the training.

Temporal-aware Discrete Visual Tokens. VQ-VAE [76] has shown promising capability in representing data as discrete tokens. VQGAN [54] further models the prior distribution of the latent space via a transformer with the GAN training. If VQGAN is directly applied onto videos, it will ignore the inner temporal coherence and treat each frame as an independent image, resulting in an unsmooth video reconstruction. Though TATS [23] attempts to handle this by making k consecutive frames altogether during VQ, it has to pre-define the constant k before training. Such constraint forbids it from representing a frame at any timestamp.

To address it with flexibility, we propose temporal-aware VQGAN (T-VQ) to inject the temporal relationship into the latent representation. We first follow VQGAN to learn the target visual tokens z_i by reconstructing a video frame v_i:

$$\begin{aligned} \mathrm{z}_i &= \mathrm{q}(\mathrm{Enc}^{\mathrm{Q}}(v_i) \mid C), \\ \hat{v}_i &= \mathrm{Dec}^{\mathrm{Q}}(\mathrm{z}_i), \\ \mathcal{L}_{\mathrm{VQ}} &= \underbrace{||\hat{v}_i - v_i||_1}_{\text{reconstrcution}} + \underbrace{||\mathrm{sg}[\mathrm{Enc}^{\mathrm{Q}}(v_i)] - C_{\mathrm{z}_i}||_2^2}_{\text{codebook}} \\ &+ \underbrace{\beta||\mathrm{sg}[C_{\mathrm{z}_i}] - \mathrm{Enc}^{\mathrm{Q}}(v_i)||_2^2}_{\text{commit}} + \underbrace{||\mathcal{F}(\hat{v}_i) - \mathcal{F}(v_i)||_1}_{\text{matching}}, \end{aligned} \tag{1}$$

where $\mathrm{Enc}^{\mathrm{Q}}$ and $\mathrm{Dec}^{\mathrm{Q}}$ are the VQ encoder and decoder. The discrete latent code z_i is acquired from the quantization operation q [54], which adopts nearest neighbor search by the trainable codebook C. We apply the straight-through estimator over the stop-gradient operation sg and use β as 0.25 [76]. We also append VGG [67] features matching $\mathcal{F}$ to stabilize the VQ loss $\mathcal{L}_{\mathrm{VQ}}$ [23]. The adversarial training between the frame quality loss $\mathcal{L}_G$ and discrimination loss $\mathcal{L}_D$ are further calculated from the discriminator D:

$$\begin{aligned} \mathcal{L}_G &= \log(1 - D(\hat{v}_i)), \\ \mathcal{L}_D &= \log(1 - D(\hat{v}_i)) + \log(D(v_i)). \end{aligned} \tag{2}$$

To inject the temporal relationship into z, T-VQ is trained with the introduced contrastive temporal reasoning:

$$\begin{aligned} o_i &= \mathrm{FC}^{\mathrm{T}}(\mathrm{z}_i, \mathrm{z}_j), \\ \mathcal{L}_{\mathrm{T}} &= \mathrm{BCELoss}(o_i, 0 \text{ if } i > j \text{ else } 1), \end{aligned} \tag{3}$$

where j is a random frame from the same video. FC^{T} is the MLP classifier, and BCELoss is the binary cross-entropy for before/after. Learning the temporal order from $\mathcal{L}_{\text{T}}$, z facilitates an implicit temporal coherence, leading to smooth video modeling. Moreover, since z represents a single image, it is flexible to support frames at arbitrary timestamps.

Generation from Masked Video. We propose the masking strategy $\mathcal{M}$ to obtain the masked videos $\overline{\mathcal{V}}$ from diverse time points. $\mathcal{M}$ masks out most video frames with the probability p and replaces each fragment as a unique [SPAN] token. For example, $\mathcal{M}$ reserves the third and the fifth frame, and masks all the others over a video length of 5:

$$\overline{\mathcal{V}} : \{[\text{S}], v_3, [\text{S}], v_5\} = \mathcal{M}(\mathcal{V} \mid p). \tag{4}$$

Our goal is to recover the missing part from $\overline{\mathcal{V}}$ and perform video completion, guided by the instruction $\mathcal{X}$. To model between the vision and language modalities, we apply our Enc^{Q} over $\overline{\mathcal{V}}$ for the discrete visual tokens $\{[\text{S}], \text{z}_3, [\text{S}], \text{z}_5\}$. We also tokenize the text $\mathcal{X}$ into word tokens $\{w_i\}_{i=1}^{L}$ with the CLIP tokenizer [56], where L is the length of $\mathcal{X}$. As in the same discrete space, MMVG can achieve cross-modal fusion by the multimodal encoder (Enc^{M}) through the self-attention mechanism as the transformer [77]:

$$\begin{aligned} f_i^w, f_j^v &= \text{LP}^w(w_i), \text{LP}^v(\text{z}_j) \\ \{h\} &= \text{Enc}^{\text{M}}([\{f^w\}, \{f^v\}]), \end{aligned} \tag{5}$$

where it obtains the features f by the linear projection (LP), and h is the hidden encoding features. We can also regard LP as the video/language embedder, which extracts the preliminary visual/linguistic features.

After encoding the language hint and the partial-missing video from Enc^{M}, our video decoder (Dec^{M}) learns to produce all frames for comprising the complete video. Dec^{M} follows the vanilla autoregressive decoder, which first conducts self-attention over the past generated tokens and then predicts the discrete visual tokens as the video frame, conditioned on the encoded features h:

$$\begin{aligned} \hat{\text{z}}_t &= \text{Dec}^{\text{M}}(\{\hat{\text{z}_1}, ..., \hat{\text{z}_{t-1}}\} \mid \{h\}), \\ \mathcal{L}_t &= \text{CELoss}(\hat{\text{z}_t}, \text{z}_t), \\ \mathcal{L}_{\text{M}} &= \sum_{t=1}^{N} \mathcal{L}_t, \end{aligned} \tag{6}$$

where z_t is the ground-truth tokens of the frame v_t in the original $\mathcal{V}$. We calculate the video decoding loss $\mathcal{L}_{\text{M}}$ by the cross-entropy (CELoss) to learn video generation as classification. Our Dec^{M} is built upon VideoSwin [45], which has shown a strong visual perception on various video understanding tasks. The 3D-shifted windows [44] consider different levels of spatial-temporal attention, and each window models blocks of video patches across T' consecutive

Algorithm 1 Mutlimodal Masked Video Generation

while TRAIN_T-VQ **do**
  $\mathcal{V} \leftarrow$ sample video
  $\text{z}_i = \text{q}(\text{Enc}^{\text{Q}}(v_i) \mid C)$
  $\hat{v_i} = \text{Dec}^{\text{Q}}(\text{z}_i)$
  $o_i = \text{FC}^{\text{T}}(\text{z}_i, \text{z}_j)$ ▷ randomly sampled frame j
  $\mathcal{L}_{\text{VQ}}, \mathcal{L}_G \leftarrow$ reconstruction, frame quality loss ▷ Eq. 1&2
  $\mathcal{L}_{\text{T}} \leftarrow$ temporal ordering loss ▷ Eq. 3
  Update T-VQ by minimizing $\mathcal{L}_{\text{VQ}}$+$\mathcal{L}_G$+$\mathcal{L}_{\text{T}}$
  $\mathcal{L}_D \leftarrow$ discrimination loss ▷ Eq. 2
  Update D by maximizing $\mathcal{L}_D$
end while

while TRAIN_MMVG **do**
  $\mathcal{V}, \mathcal{X}, p \leftarrow$ sample video/instruction/probability
  $\overline{\mathcal{V}}$: $\{v_a, [\text{S}], v_b, ...\} = \mathcal{M}(\mathcal{V} \mid p)$ ▷ diverse guided frames
  $\{\text{z}_a, [\text{S}], \text{z}_b, ...\}, \{w\} = \text{Enc}^{\text{Q}}(\overline{\mathcal{V}})$, Tokenizer($\mathcal{X}$)
  $\{h\} = \text{Enc}^{\text{M}}([\{w\}, \{\text{z}_a, [\text{S}], \text{z}_b, ...\}])$
  for $t \leftarrow 1$ to N **do**
    $\hat{\text{z}_t} = \text{Dec}^{\text{M}}(\{\text{z}_1, ..., \text{z}_{t-1}\} \mid \{h\})$ ▷ teacher-forcing
    $\mathcal{L}_t \leftarrow$ video decoding loss ▷ Eq. 6
  end for
  $\hat{\mathcal{V}} = \text{Dec}^{\text{Q}}(\{\hat{\text{z}}_{t=1}^{N}\})$
  $\mathcal{L}_{\text{M}} = \sum_{t=1}^{N} \mathcal{L}_t$
  Update MMVG by minimizing $\mathcal{L}_{\text{M}}$
  $p \leftarrow$ update masking probability ▷ Eq. 8
end while

frames. To ensure the same dimension for video generation in Dec^{M}, we remove the temporal down-sampling layer. In the end, we can utilize Dec^{Q} to reconstruct all the frames as our completed videos $\hat{\mathcal{V}}$:

$$\hat{\mathcal{V}} = \text{Dec}^{\text{Q}}(\{\hat{\text{z}}\}_{t=1}^{N}). \tag{7}$$

By varying the masking conditions through $\mathcal{M}$, MMVG learns how to complete a video from partial frames $\overline{\mathcal{V}}$ at arbitrary time points with the text, which overcomes the limitation of chronological guidance. To make $\mathcal{M}$ more effective, we apply an adaptive probability p instead of random sampling every time. Each video $\mathcal{V}$ keeps its own p, and all frames are equally initialized in the beginning. Based on the prediction error, we adjust the masking probability p_t of the t-th frame:

$$p_t = p_t + \alpha((\frac{\mathcal{L}_t}{\mathcal{L}_M} \sum p) - p_t), \tag{8}$$

where α is the adjusting rate. A larger video decoding loss $\mathcal{L}_t$ indicates that the t-th frame is more difficult to recover. MMVG learns more from those challenging cases and can bring better generative quality for video completion.

Unifying TVC during Inference. After training with text and partial-missing video, MMVG learns to perform video completion over [SPAN] tokens. Then for inference, Enc^{M} takes the following as its input to support different tasks:

Method	Text	TVPrediction						TVRewind						TVInfilling					
		Kitchen		Flintstones		MUGEN		Kitchen		Flintstones		MUGEN		Kitchen		Flintstones		MUGEN	
		FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑	FVD↓	RCS↑
FILM [60]	✗	-	-	-	-	-	-	-	-	-	-	-	-	250.2	56.1	352.7	51.4	538.8	5.9
VideoMAE [73]	✗	328.9	47.6	317.5	55.6	548.7	7.0	365.9	48.2	335.5	55.9	545.2	7.1	246.9	54.7	211.5	60.6	494.9	7.8
TATS [23]	✗	106.9	64.4	127.5	60.3	376.5	7.1	107.7	62.7	127.6	60.2	350.8	7.2	71.5	72.7	119.5	66.7	328.2	8.4
MMVG^{U}	✗	105.6	63.3	124.8	60.5	374.5	7.2	109.8	62.6	124.3	59.7	356.4	7.0	71.5	73.4	121.8	66.3	328.4	7.8
MMVG^{S}	✗	**103.8**	**64.5**	**123.8**	**60.8**	**369.4**	**7.3**	**105.9**	**63.6**	**123.8**	**60.5**	**347.8**	**7.2**	**68.5**	**73.6**	**118.5**	**67.9**	**324.3**	**8.4**
TATS [23]	✓	87.2	66.3	115.9	70.6	90.1	67.9	89.8	63.3	116.3	70.4	89.8	68.7	57.4	77.6	95.8	78.2	58.9	73.6
MMVG^{U}	✓	80.2	68.4	108.2	72.9	84.8	70.2	83.2	66.9	113.2	71.6	93.1	68.4	59.8	77.8	92.8	78.3	59.2	73.2
MMVG^{S}	✓	**75.6**	**68.8**	**106.3**	**73.7**	**83.3**	**71.1**	**79.7**	**68.1**	**107.2**	**72.9**	**88.7**	**70.0**	**56.0**	**78.1**	**91.6**	**79.6**	**57.2**	**74.1**

Table 1. Results of TVC, including prediction, rewind, and infilling, on Kitchen [13], Flintstones [26], and MUGEN [29]. TATS [23] requires specific training to support different tasks. We further train the unified MMVG^{U} for each specific task as MMVG^{S}.

- TVPrediction: $[\{w\}, \{\mathrm{z}_1, [\mathrm{SPAN}]\}]$
- TVRewind: $[\{w\}, \{[\mathrm{SPAN}], \mathrm{z}_N\}]$
- TVInfilling: $[\{w\}, \{\mathrm{z}_1, [\mathrm{SPAN}], \mathrm{z}_N\}]$

In this way, a single trained MMVG can unify all TVC tasks without the specific downstream fine-tuning.

3.3. Learning of MMVG

Algo. 1 illustrates the learning process of our proposed MMVG for TVC. We first train T-VQ over video frames for discrete visual tokens with contrastive temporal reasoning. Specifically, we minimize the VQ reconstruction loss $\mathcal{L}_{\mathrm{VQ}}$ and frame quality loss $\mathcal{L}_G$ along with our temporal ordering loss $\mathcal{L}_{\mathrm{T}}$ to optimize T-VQ. At the same time, we also update the discriminator D via the standard adversarial training by maximizing the discrimination loss $\mathcal{L}_D$. For video completion, the masking strategy $\mathcal{M}$ masks the video frames with the probability p and then acquires guided frames from diverse time points. MMVG regards text and partial-missing video by $\mathrm{Enc}^{\mathrm{M}}$ for cross-modal fusion, and $\mathrm{Dec}^{\mathrm{M}}$ further predicts the visual tokens of frames autoregressively as the complete video. As a sequential generation process, we apply the teacher-forcing trick. Instead of our predicted $\hat{\mathrm{z}}$, the ground-truth z from the previous timestamp is fed to stabilize the training. Each video decoding loss $\mathcal{L}_t$ at timestamp t is summed up as $\mathcal{L}_{\mathrm{M}}$ to optimize MMVG. According to $\mathcal{L}_t$, we update p for effective masking probability. The entire optimization object can be summarized as two phases:

$$\begin{aligned} \text{T-VQ:} &\min_{\mathrm{Enc}^{\mathrm{Q}},\mathrm{Dec}^{\mathrm{Q}},C,\mathrm{FC}^{\mathrm{T}}} \max_{D} \mathcal{L}_{\mathrm{VQ}} + \mathcal{L}_G + \mathcal{L}_D + \mathcal{L}_{\mathrm{T}} \\ \text{MMVG:} &\min_{\mathrm{Enc}^{\mathrm{M}},\mathrm{Dec}^{\mathrm{M}}} \mathcal{L}_{\mathrm{M}} \end{aligned} \tag{9}$$

4. Experiments

4.1. Experimental Setup

Datasets. As a new task, we consider diverse video scenes with natural instructions for TVC. **Kitchen** [13] records 22K egocentric videos about kitchen activity, which have different lengths (4-16 frames) with narrations. **Flintstones** [26] contains 25K animation videos (15 frames) from *The Flintstones*, where each video description includes the characters and their behavior. **MUGEN** [29] is built from agents playing CoinRun [11], which consists of 375K gaming videos (16 frames) with detailed text annotations. All videos in these three datasets are resized into 128x128. An overview is shown in Table 2 and Fig. 1(b). Since MMVG can unify various tasks, we also evaluate video generation/prediction on widely-used **UCF-101** [69] and **BAIR** [18], video infilling on UCF-101 following RaMViD [39], and text-to-video generation on **MSRVTT** [89].

Dataset	Train / Val	#Frame	#Word	FPS
Kitchen [13]	16,695 / 5,804	8.3	2.8	6
Flintstones [26]	22,666 / 2,518	15	16.5	5
MUGEN [29]	362,239 / 12,848	16	20.6	5

Table 2. The statistics of our used datsets to evaluate TVC.

Evaluation Metrics. We apply the following metrics to evaluate TVC results: 1) **FVD** [75] computes the video features [6] distance to the ground truth; 2) **RCS** [86] is the relative visual-text similarity to the instruction, compared to the ground-truth video. We fine-tune the CLIP model [56] on each dataset and adapt it to the video scene for a more precise alignment. Apart from automatic metrics, we also conduct a human evaluation from aspects of video quality, instruction relevance, and ground-truth similarity. We sample 75 TVP results for each task and adopt MTurk[1] to rank over baselines and our MMVG. To avoid the potential ranking bias, we hire 3 MTurkers for each sample of prediction, rewind, and infilling tasks.

Implementation Detail. T-VQ contains ResBlocks [31] as the visual auto-encoder ($\mathrm{Enc}^{\mathrm{Q}}$ and $\mathrm{Dec}^{\mathrm{Q}}$). The discriminator D follows a similar architecture to $\mathrm{Enc}^{\mathrm{Q}}$. For the vector quantization, we use a patch size 16, where a 128x128 video frame transforms into 8x8 discrete visual tokens. There are 1024 vocabularies in the codebook C, and the hidden embedding size is 256. We adopt batch size 32 with a learning rate of 4.5e-6 to optimize T-VQ by Adam [42]. MMVG is built in an encoder-decoder manner, where $\mathrm{Enc}^{\mathrm{M}}$ is a trans-

[1]Amazon MTurk: https://www.mturk.com. Our studies have been cleared by the human subject committee as an IRB-exempt protocol.

Method	Kitchen	Flintstones	MUGEN
VideoDiff [34]	138.6	206.4	410.7
MCVD [79]	119.9	183.8	400.2
TATS [23]	115.5	157.5	386.4
MMVG	**109.1**	**127.6**	**368.6**

Table 3. **FVD** results of **video generation** on our TVC datasets.

Method	UCF-101 IS↑	UCF-101 FVD↓
VideoGPT [94]	24.7	-
DIGAN [96]	32.7	577
VideoDiff [34]	57.0	-
TATS [23]	57.6	420
MMVG	**58.3**	**395**

Table 4. Results of **video generation** on UCF-101 [69].

Method	BAIR FVD↓
VideoGPT [94]	103.3
MaskViT [25]	93.6
MCVD [79]	89.5
TATS [23]	88.6
MMVG	**85.2**

Table 5. Results of **video prediction** on BAIR [18].

former with 24 layers, 16 attention heads, and hidden embedding size 1024. Dec^{M} adopts a similar setting with temporal window size 3 in VideoSwin. The initial sample rate p of the masking strategy $\mathcal{M}$ is 0.9 with an adjusting rate α as 0.1. We optimize MMVG through the mixed precision [48] with batch size 4 by Adam. The learning rate is also 4.5e-6. All experiments are implemented in PyTorch [53] and done on 8 NVIDIA A100 GPUs.

4.2. Main Results

Table 1 shows the results of all text-guided prediction, rewind, and infilling for TVC. VideoMAE [73] is built upon MAE [30] and reconstructs the missing video cubes, which performs TVC by masking all video frames except the first or the last (or both). TATS [23], the SOTA on video generation, also produces videos as discrete visual tokens. Since TATS can only consider the past through the autoregressive transformer, it requires specific training for each task. We have MMVG^{U} as the unified model that can support all TVC tasks with a single training and MMVG^{S} to further train for each prediction, rewind, and infilling. We treat TATS as our main baseline[2] and study the importance of guided text.

TVPrediction. VideoMAE attempts to produce all frames simultaneously, which is difficult to maintain video temporal consistency, resulting in a high 328.9 FVD on Kitchen. TATS is inherently designed for prediction as it generates the frames one after one. However, our unified MMVG^{U} performs better than TATS on all datasets (*e.g.*, lower 105.6 and 124.8 FVD on Kitchen and Flintstones). These results support that learning from diverse time points will not hurt the prediction from the past. In contrast, our masking strategy can bring superior temporal coherence. MMVG^{S} further improves itself through training prediction as completion from the head. However, there are too many possible outcomes from just the beginning, where the predicted results may not meet the expectation (*e.g.*, a high 370 FVD on MUGEN). The instruction as guidance makes it related to the expected ground-truth result. We can let MUGEN run, jump, or collect coins as the textual descriptions to achieve more controllability, leading to a noticeable improvement (*e.g.*, a lower 84.8 FVD by MMVG^{U}). The higher 70.2 RCS also shows that our MMVG can produce MUGEN videos that confirm with the instruction. Although the model may try to imagine the animation or the kitchen activity, the language hint can provide a clear goal to anticipate. Likewise, MMVG^{U} with text surpasses TATS, even though it is not designed for prediction only. The specific trained MMVG^{S} benefits the unified model for further improvement.

TVRewind. Rewind from the last allows the model to imagine what happened along with a suitable opening. In addition, the objects may not display on the last frame (*e.g.*, the spoons and forks for "*close drawer*"), which makes it more challenging to complete. Similar to prediction, VideoMAE cannot have feasible rewind results. Language is still essential to remind the past and establish an adequate beginning, where we can find a significant performance gap between with and without text (*e.g.*, 90 *vs.* 350 FVD on MUGEN). Our unified MMVG^{U} achieves comparable results to TATS and even outperforms on Kitchen and Flintstones (*e.g.*, higher 66.9 and 71.6 RCS). With the learning of completion from partial frames, autoregressive model can still accomplish video rewind without specific training. If following TATS design to train MMVG^{U} for rewind, MMVG^{S} gains more improvement and utterly surpasses it.

TVInfilling. We consider the additional FILM [60] for infilling, which performs video interpolation with in-between motion. Despite synthesizing intermediate frames between the first and the last, the visual dynamics are changing too rapidly to handle, resulting in a higher FVD. With guidance from the head and tail, we find a noticeable improvement even without instruction (*e.g.*, lower FVDs on Kitchen), which is helpful in temporal video modeling. To capture the expected missing event, we still require the language hint for more controllability. Our unified MMVG^{U} achieves comparable performance to TATS again, which is specifically trained for the infilling task. It shows that completion from partial frames at different time points still helps, and MMVG^{S} further outperforms on TVInfilling.

4.3. Additional Study

Video Generation/Prediction. We further evaluate the classic video generation and prediction tasks. Table 3 shows FVD scores of unconditional video generation on our TVC datasets. Note that only videos but no texts are used in these experiments. Both VideoDiff [34] and MCVD [79] are built

[2]Since we cannot receive feasible results after training diffusion methods for our TVC, we evaluate unconditional video generation in Sec. 4.3. We use this repo (https://github.com/lucidrains/video-diffusion-pytorch) as VideoDiff and the official repo as MCVD.

Method	UCF-101 (FVD↓) K=+1	+2	+5	±1	±2
RaMViD [39]	349.7	300.6	260.5	215.4	162.5
MMVG	**316.3**	**258.5**	**194.6**	**183.2**	**120.3**

Table 6. Results of **video prediction** and **infilling** on UCF-101.

upon denoising diffusion [33], where MCVD also considers different partial frames during training. The results first indicate that the vanilla token-based method is superior to the diffusion-based model (TATS *vs.* VideoDiff) for video generation. In addition, MMVG, with the masking strategy that learns the visual guidance from diverse time points, further boosts the performance (the lowest 127.6, 109.1, and 368.6 FVD on Flintstones, Kitchen, and MUGEN, respectively).

We also evaluate MMVG on UCF-101 [69], which is challenging to generate natural human videos. Table 4 supports that our MMVG can produce videos with higher visual similarity (a higher 58.3 IS) and temporal alignment (a lower 395 FVD) to the ground truth. For video prediction, we apply the widely-used BAIR [18] in Table 5, where the model has to anticipate how a robot pushes objects from the given first frame. MMVG again surpasses TATS with the lowest 85.2 FVD. Although both generation and prediction are generating video frames chronologically, the ability to recover arbitrary missing frames for video completion empowers MMVG with a stronger temporal coherence, leading to better generative video quality.

Video Infilling. We follow RaMViD [39] to evaluate video infilling on UCF-101. We consider various guidance settings K in Table 6. For example, K=+1 means given the first frame, and K=±2 should provide the first and last two frames. For prediction, MMVG outperforms RaMViD on all K, and the performance gap gets even larger when more guided frames are accessible (*e.g.*, 33.4 on K=+1 and 65.9 on K=+5). A similar result can be found for infilling, where MMVG can make the lowest 120.3 FVD on K=±2. Despite having a similar masking strategy, it shows that generating frames one after one still brings superior results. MMVG allows autoregressive models to condition on visual hints from any time point, which produces more similar videos to the ground truth when infilling between the head and tail.

Text-to-Video Generation. Being a multimodal generative model, MMVG supports text-to-video generation. To compare with those large-scale methods, we pre-train MMVG using WebVid [4], which contains 2.5M text-video pairs. We adopt the masking strategy to treat the pre-training as video completion. MMVG outperforms GODIVA [86] and NUWA [87] without access the MSRVTT [89] data in Table 7. Surprisingly, MMVG can generate videos that are more related to the texts (a higher 0.2644 CLIP-S [86]) than CogVideo [35], even though using twice less data. This result encourages the effectiveness of completion from partial

Method	Pre-training	Zero-shot	MSRVTT FID↓	CLIP-S↑
GODIVA [86]	136M	✗	-	0.2402
NUWA [87]	3.9M	✗	47.7	0.2439
CogVideo [35]	5.4M	✓	23.6	0.2631
Make-A-Video [68]	20M	✓	13.2	0.3049
MMVG	2.5M	✓	**23.4**	**0.2644**
TATS [23]	-	✗	63.2	0.2326
MMVG	-	✗	**60.6**	**0.2385**

Table 7. Results of **text-to-video generation** on MSRVTT [89]. We gray out methods that use significantly more pre-training data.

Method	Text	Kitchen Q.	T.	GT	Flintstones Q.	T.	GT	MUGEN Q.	T.	GT
MMVG	✗	1.99	1.81	1.82	1.73	1.66	1.62	**2.03**	1.56	1.55
TATS [23]	✓	1.97	2.07	2.03	2.07	2.12	2.17	1.94	2.11	2.19
MMVG	✓	**2.04**	**2.12**	**2.15**	**2.20**	**2.22**	**2.21**	2.03	**2.33**	**2.26**

Table 8. Human evaluation for TVP with aspects of video quality (Q.), instruction relevance (T.), and ground-truth similarity (GT).

frames. For a fair comparison without additional data, we directly train on MSRVTT. MMVG still outperforms TATS, which shows that text-to-video generation can be improved through learning from video completion as well.

Human Evaluation. We study the video quality (Q.), the relevance to the instructions (T.), and the similarity to the ground-truth video (GT) of the produced videos from the human aspect. The results in Table 8 are calculated as the mean ranking score (from 1 to 3, the higher is better) of each method for TVPrediction. MMVG without text even generates higher quality videos than TATS with text on Kitchen and MUGEN, where completion from partial frames benefits the temporal coherence of generative video modeling. While, the lowest ground-truth similarity illustrates that language guidance is crucial for controllability. With instruction, MMVG anticipates the future as the text (the highest T.) and generates videos that meet the ground truth (the highest GT), achieving the best overall performance.

Qualitative Results. Fig. 3 illustrates the keyframes of the generated examples of three datasets. Thanks to the learning of completion from partial frames at diverse time points, a single trained MMVG can support all TVC tasks. For prediction, MMVG makes Fred "*turn his head*" or MUGEN "*jumps over the gear*" from the guided text. MMVG further recovers the missing spoons and forks for "*close drawer*" from the last frame in a more challenging rewind scenario. MMVG infills the missing event described in language such as "*stand up for dancing*", "*walk across the kitchen*", "*jump onto the stage*" from the head and tail.

From the same visual guidance, MMVG can lead TVC results using different texts, achieving controllable video completion. For example, we can let MUGEN "*jump down*

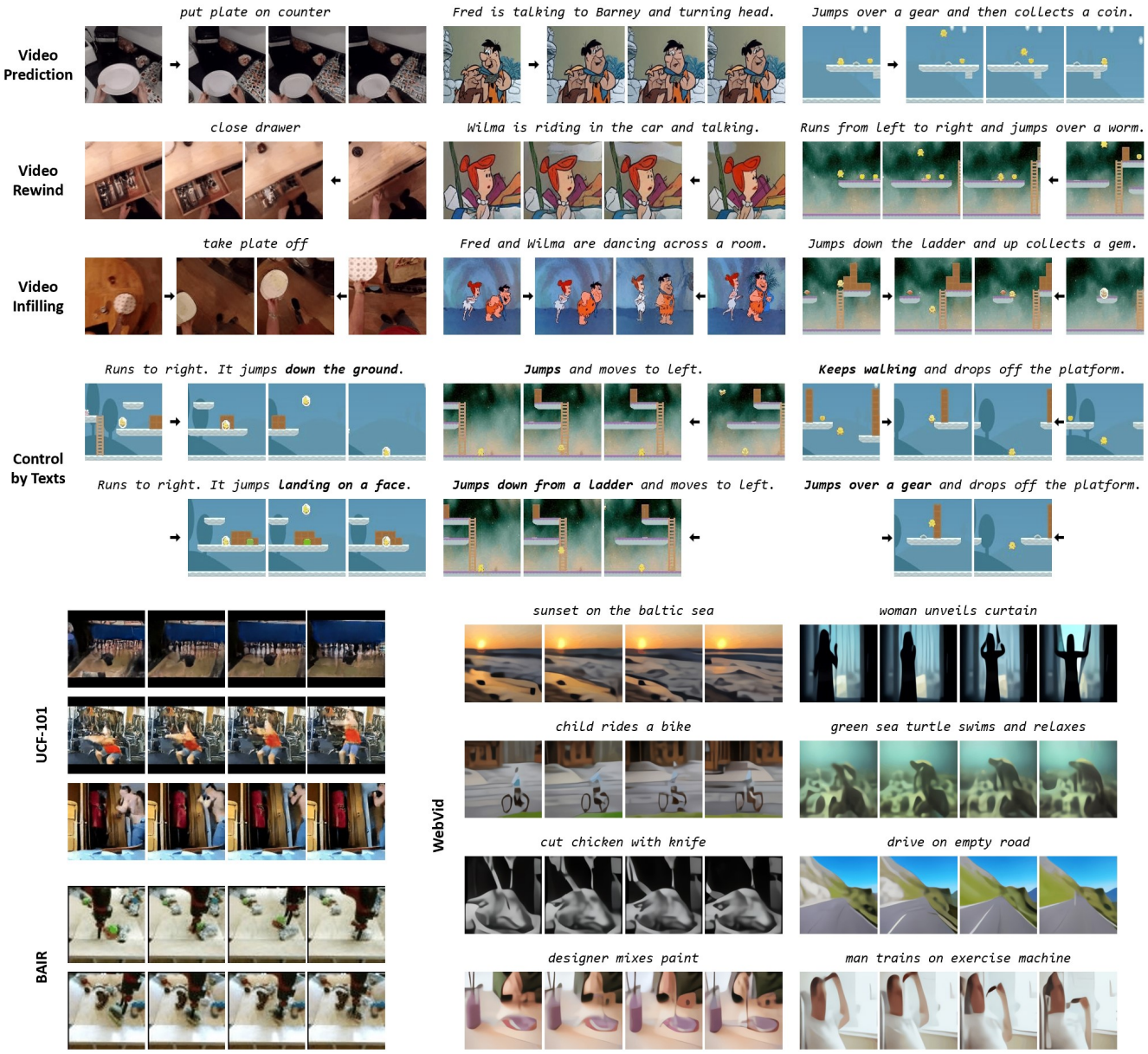

Figure 3. Qualitative examples of TVC on Kitchen [13], Flintstones [26], and MUGEN [29]. We also illustrate video generation on UCF-101 [69], video prediction on BAIR [18], and text-to-video prediction on WebVid [4].

the ground" or "*land on a face*", starting from the same beginning. We can also control the behavior as "*keep walking*" or "jump over a gear" to recover the missing middle event. Furthermore, MMVG also carries out unconditional video generation with smooth temporal coherence. We can use language to produce natural dynamics in diverse scenes, such as "*sunset on the sea*", "*green sea turtle swins*", or a close look of "*cut chicken*". The presented videos indicate that our method not only unifies all TVC tasks but also performs the classical video generation/text-to-video well.

5. Conclusion

We introduce a novel task of text-guided video completion (TVC) that performs video completion from the first, last, or both frame(s) controlled by language. We present Multimodal Masked Video Generation (MMVG) with an effective masking strategy to learn the visual guidance from any time point. By varying the masking conditions, MMVG addresses all prediction, rewind, and infilling tasks within one model. Experiments on various video scenes show that our MMVG effectively addresses TVC as well as generative video modeling. We believe TVC can help advance a new field toward vision-and-language research.

References

[1] Dinesh Acharya, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards High Resolution Video Generation with Progressive Growing of Sliced Wasserstein GANs. In *arXiv:1810.02419*, 2018. 2

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic Variational Video Prediction. In *International Conference for Learning Representations (ICLR)*, 2018. 2

[3] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. FitVid: Overfitting in Pixel-Level Video Prediction. In *arXiv:2106.13195*, 2021. 2

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 7, 8

[5] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2LIVE: Text-Driven Layered Image and Video Editing. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[6] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[7] Ya-Liang Chang, Zhe Liu Yu, and Winston Hsu. VORNet: Spatio-temporally Consistent Video Inpainting for Object Removal. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 2

[8] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent Online Video Style Transfer. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[9] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Jonathan Warrell, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy Mitra, and Philip Torr. ImageSpirit: Verbal Guided Image Parsing. In *Conference on Human Factors in Computing Systems (CHI)*, 2013. 2

[10] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial Video Generation on Complex Datasets. In *arXiv:1907.06571*, 2019. 2

[11] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying Generalization in Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2019. 5

[12] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 5, 8

[14] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary Video Style Transfer via Multi-Channel Correlation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[15] Emily Denton and Rob Fergus. Stochastic Video Generation with a Learned Prior. In *International Conference on Machine Learning (ICML)*, 2018. 1

[16] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *arXiv:2105.05233*, 2021. 2

[17] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2

[18] Frederik Ebert, Chelsea Finn, Alex X. Lee, and Sergey Levine. Self-Supervised Visual Planning with Temporal Skip Connections. In *Conference on Robot Learning (CoRL)*, 2017. 2, 5, 6, 7, 8

[19] Alaaeldin El-Nouby, Shikhar Sharma, Hannes Schulz, Devon Hjelm, Layla El Asri, Samira Ebrahimi Kahou, Yoshua Bengio, and Graham W.Taylor. Tell, Draw, and Repeat: Generating and Modifying Images Based on Continual Linguistic Instruction. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[20] Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. SSCR: Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[21] Tsu-Jui Fu, Xin Eric Wang, Scott Grafton, Miguel Eckstein, and William Yang Wang. M^3L: Language-based Video Editing via Multi-Modal Multi-Level Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[22] Tsu-Jui Fu, Xin Eric Wang, and William Yang Wang. Language-Driven Artistic Style Transfer. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[23] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 5, 6, 7

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015. 2

[25] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. MaskViT: Masked Visual Pre-Training for Video Prediction. In *arXiv:2206.11894*, 2022. 2, 6

[26] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine This! Scripts to Compositions to Videos. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 5, 8

[27] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[28] Zekun Hao, Xun Huang, and Serge Belongie. Controllable Video Generation with Sparse Trajectories. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[29] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Qiyuan Hu, and Devi Parikh. MUGEN: A Playground for Video-Audio-Text Multimodal Understanding and GENeration. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 5, 8

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *International Conference on Computer Vision (ICCV)*, 2015. 5

[32] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models. In *arXiv:2210.02303*, 2022. 2

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2, 7

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video Diffusion Models. In *arXiv:2204.03458*, 2022. 2, 6

[35] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In *arXiv:2205.15868*, 2022. 2, 7

[36] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Learning to Decompose and Disentangle Representations for Video Prediction. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 1

[37] Qiyang Hu, Adrian Waelchli, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Video Synthesis from a Single Image and Motion Stroke. In *arXiv:1812.01874*, 2018. 1

[38] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make It Move: Controllable Image-to-Video Generation with Text Descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[39] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion Models for Video Prediction and Infilling. In *arXiv:2206.07696*, 2022. 1, 2, 5, 7

[40] Nal Kalchbrenner, Aaron van den Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video Pixel Networks. In *International Conference on Machine Learning (ICML)*, 2017. 1

[41] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Deep Video Inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[42] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference for Learning Representations (ICLR)*, 2015. 5

[43] Yitong Li, Martin Renqiang Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video Generation From Text. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *International Conference on Computer Vision (ICCV)*, 2021. 4

[45] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[46] Tanya Marwah, Gaurav Mittal, and Vineeth N. Balasubramanian. Attentive Semantic Video Generation using Captions. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[47] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep Multi-Scale Video Prediction Beyond Mean Square Error. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[48] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed Precision Training. In *International Conference for Learning Representations (ICLR)*, 2018. 6

[49] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. In *arXiv:1411.1784*, 2014. 2

[50] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & Play Generative Networks: Conditional Iterative Generation of Images in Latent Space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[51] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning (ICML)*, 2022. 2

[52] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To Create What You Tell: Generating Videos from Captions. In *International Conference on Multimedia (MM)*, 2017. 2

[53] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *International Conference on Learning Representations Workshop (ICLRW)*, 2017. 6

[54] Björn Ommer Patrick Esser, Robin Rombach. Taming Transformers for High-Resolution Image Synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[55] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. MirrorGAN: Learning Text-to-image Generation by Redescription. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Interna-*

tional Conference on Machine Learning (ICML), 2021. 4, 5

[57] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent Video Transformer. In *arXiv:2006.10704*, 2020. 2

[58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125*, 2022. 2

[59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, 2021. 2

[60] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. FILM: Frame Interpolation for Large Motion. In *European Conference on Computer Vision (ECCV)*, 2022. 5, 6

[61] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative Adversarial Text to Image Synthesis. In *International Conference on Machine Learning (ICML)*, 2016. 2

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *arXiv:2205.11487*, 2022. 2

[64] Masaki Saito and Shunta Saito. TGANv2: Efficient Training of Large Models for Video Generation with Multiple Subsampling Layers. In *arXiv:1811.09245*, 2018. 1

[65] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *arXiv:2111.02114*, 2021. 2

[66] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meetings of the Association for Computational Linguistics (ACL)*, 2018. 2

[67] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[68] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *arXiv:2209.14792*, 2022. 2, 7

[69] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In *arXiv:1212.0402*, 2012. 2, 5, 6, 7, 8

[70] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. In *International Conference on Machine Learning (ICML)*, 2015. 1

[71] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2Scene: Generating Compositional Scenes from Textual Descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[72] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. A Good Image Generator Is What You Need for High-Resolution Video Synthesis. In *International Conference for Learning Representations (ICLR)*, 2021. 2

[73] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[74] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[75] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, and Sylvain Gelly Marcin Michalski. Towards Accurate Generative Models of Video: A New Metric & Challenges. In *International Conference on Learning Representations Workshop (ICLRW)*, 2019. 5

[76] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4

[78] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable Length Video Generation From Open Domain Textual Description. In *arXiv:2210.02399*, 2022. 2

[79] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked Conditional Video Diffusion for Prediction, Generation, and Interpolation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 6

[80] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2016. 2

[81] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2

[82] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2

[83] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A Large-Scale,

High-Quality Multilingual Dataset for Video-and-Language Research. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[84] Dirk Weissenborn, Oscar Tackstrom, and Jakob Uszkoreit. Scaling Autoregressive Video Models. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[85] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling Autoregressive Video Models. In *International Conference for Learning Representations (ICLR)*, 2020. 2

[86] 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. In *arXiv:2104.14806*, 2021. 2, 5, 7

[87] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 7

[88] Xide Xia, Tianfan Xue, Wei-Sheng Lai, Zheng Sun, Abby Chang, Brian Kulis, and Jiawen Chen. Real-Time Localized Photorealistic Video Style Transfer. In *Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[89] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 7

[90] Qiangeng Xu, Hanwang Zhang, Weiyue Wang, Peter N. Belhumeur, and Ulrich Neumann. Stochastic Dynamics for Video Infilling. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[91] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep Flow-Guided Video Inpainting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[92] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[93] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally Consistent Semantic Video Editing. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[94] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video Generation using VQ-VAE and Transformers. In *arXiv:2104.10157*, 2021. 2, 6

[95] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. In *arXiv:2206.10789*, 2022. 2

[96] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating Videos with Dynamics-aware Implicit Generative Adversarial Networks. In *International Conference for Learning Representations (ICLR)*, 2022. 2, 6

[97] 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 *European Conference on Computer Vision (ECCV)*, 2020. 1