Show Me What and Tell Me How: Video Synthesis via Multimodal Conditioning

Ligong Han[1,2*] Jian Ren[1] Hsin-Ying Lee[1] Francesco Barbieri[1]
Kyle Olszewski[1] Shervin Minaee[1] Dimitris Metaxas[2] Sergey Tulyakov[1]
[1]Snap Inc. [2]Rutgers University

Abstract

Most methods for conditional video synthesis use a single modality as the condition. This comes with major limitations. For example, it is problematic for a model conditioned on an image to generate a specific motion trajectory desired by the user since there is no means to provide motion information. Conversely, language information can describe the desired motion, while not precisely defining the content of the video. This work presents a multimodal video generation framework that benefits from text and images provided jointly or separately. We leverage the recent progress in quantized representations for videos and apply a bidirectional transformer with multiple modalities as inputs to predict a discrete video representation. To improve video quality and consistency, we propose a new video token trained with self-learning and an improved mask-prediction algorithm for sampling video tokens. We introduce text augmentation to improve the robustness of the textual representation and diversity of generated videos. Our framework can incorporate various visual modalities, such as segmentation masks, drawings, and partially occluded images. It can generate much longer sequences than the one used for training. In addition, our model can extract visual information as suggested by the text prompt, e.g.*, "an object in image one is moving northeast", and generate corresponding videos. We run evaluations on three public datasets and a newly collected dataset labeled with facial attributes, achieving state-of-the-art generation results on all four*[1].

1. Introduction

Generic video synthesis methods generate videos by sampling from a random distribution [56, 57]. To get more control over the generated content, conditional video synthesis works utilize input signals, such as images [9, 19], text or language [7, 30], and action classes [62]. This enables synthesized videos containing the desired objects as specified by visual information or desired actions as specified by textual information.

Existing works on conditional video generation use only one of the possible control signals as inputs [8, 30]. This limits the flexibility and quality of the generative process. For example, given a screenplay we could potentially generate several movies, depending on the decisions of the director, set designer, and visual effect artist. In a similar way, a video generation model conditioned with a text prompt should be primed with different visual inputs. Additionally, a generative video model conditioned on a given image should be able to learn to generate various plausible videos, which can be defined from various natural language instructions. For example, to generate object-centric videos with objects moving [70], the motion can be easily defined through a text prompt, *e.g.*, "moving in a zig-zag way," while the objects can be defined by visual inputs. Thus, an interesting yet challenging question arises: *Can we learn a video generation model that can support such behavior?*

We tackle the question in this work and propose a new video synthesis model supporting diverse, multimodal conditioning signals. Our method consists of two phases. The *first* phase obtains discrete representations from images. We employ an autoencoder with a quantized bottleneck, inspired by the recent success of two-stage image generation using quantized feature representations [21, 38, 47, 76]. The *second* phase learns to generate video representations that are conditioned on the input modalities, which can then be decoded into videos using the decoder from the first stage. We leverage a bidirectional transformer, *i.e.*, BERT [17], trained with a masked sequence modeling task, that uses tokens from multimodal samples and predicts the latent representation for videos. Building such a framework requires solving several challenging problems. First, video consistency is a common problem among video generation methods. Second, it is necessary to ensure that the correct textual information is learned. Third, training a transformer for image synthesis is computationally demanding [12], an issue that is even more severe in the time domain, as a longer sequence of tokens needs to be learned. To solve these challenges, we propose the following contributions:

*Work done during an internship at Snap Inc.
[1]Code: https://github.com/snap-research/MMVID and Webpage.

- We introduce a bidirectional transformer with several new techniques to improve video generation: For training, we propose the video token `VID`, which is trained via self-learning and video attention, to model temporal consistency; For inference, we improve mask-predict to generate videos with improved quality.
- We introduce text augmentation, including text dropout and pretrained language models for extracting textual embeddings, to generate diverse videos that are correlated with the provided text.
- We explore long sequence synthesis with the transformer model to generate sequences with lengths that are much longer than the one used for training (Fig. 5).

We name our framework **MMVID** and show that a **M**ulti**M**odal **VID**eo generator can enable various applications. The user can show *what* to generate using visual modalities and tell *how* to generate with language. We explore two settings for multimodal video generation. The first involves *independent* multimodalities, such that there is no relationship between textual and visual controls (Fig. 3a and Fig. 4). The second one targets *dependent* multimodal generation, where we use text to obtain certain attributes from given visual controls (Fig. 3b and Fig. 4). The latter case allows for more potential applications, in which language is not able to accurately describe certain image content that the user seeks to generate, but images can efficiently define such content. We also show our model can use diverse visual information, including segmentation masks, drawings, and partially observed images (Fig. 4).

To validate our approach extensively, we conduct experiments on *four* datasets. In addition to three public datasets, we collect a new dataset, named Multimodal VoxCeleb, that includes $19,522$ videos from VoxCeleb [37] with 36 manually labeled facial attributes.

2. Related Works

Video Generation. For simplicity, previous works on video generation can be categorized into unconditional and conditional generation, where most of them apply similar training strategies: adversarial training with image and video discriminators [14, 24, 50]. Research on *unconditional* video generation studies how to synthesize diverse videos with input latent content or motion noise [1, 28, 49, 57, 62, 66, 75]. Recent efforts in this direction have achieved high resolution and high quality generation results for images and videos [14, 29, 56]. On the other hand, *conditional* video generation utilizes given visual or textual information for video synthesis [9, 19, 39, 48, 54, 64, 67]. For example, the task of video prediction uses the provided first image or a few images to generate a sequence of frames [3–6, 11, 15, 16, 26, 60, 61, 63]. Similarly, text-to-video generation applies the conditional signal from text, captions, or natural language descriptions [31, 36, 40]. TFGAN [7] proposes a multi-scale text-filter conditioning scheme for discriminators. TiVGAN [30] proposes to generate a single image from text and synthesizes consecutive frames through further stages. In this work, we study conditional video synthesis. However, we differ from existing methods since we address a more challenging problem: multimodal video generation. Instead of using a single modality, such as textual guidance, we show how multiple modalities can be input within a single framework for video generation. With multimodal controls, *i.e.*, textual and visual inputs, we further enhance two settings for video generation: independent and dependent multimodal inputs, in which various applications can be developed.

Transformers for Video Generation. Transformer-based networks have shown promising and often superior performance not only in natural languages processing tasks [10, 43, 59], but also in computer vision related efforts [12, 20, 27, 41, 42]. Recent works prvoide promising results on conditional image generation [13, 21], text-to-image generation [18, 45], video generation [44, 68, 70, 72], and text-to-video synthesis [69] using transformers. Unlike existing transformer-based video generation works that focus on autoregressive training, we apply a non-autoregressive generation pipeline with a bidirectional transformer [22, 23, 25, 35, 65]. Our work is inspired by M6-UFC [76], which utilizes the non-autoregressive training for multimodal image generation and produces more diverse image generation with higher quality. Building upon M6-UFC, we further introduce training techniques for multimodal video synthesis.

3. Methods

Our framework for multimodal video generation is a two-stage image generation method. It uses discrete feature representations [21, 38, 47, 76]. During the *first* stage we train an autoencoder (with encoder E and decoder D) that has the same architecture as the one from VQGAN [21] to obtain a quantized representation for images. Given a real video clip defined as $\mathbf{v} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T\}$ with $\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$, we get a quantized representation of the video defined as $\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_T\}$, where $\mathbf{z}_t = q(E(\mathbf{x}_t)) \in \mathbb{N}_1^{h\times w}$. $q(\cdot)$ denotes the quantization operation and $\mathbb{N}_1$ indicates a set of positive integers.

During the *second* stage we learn a bidirectional transformer for modeling the correlation between multimodal controls and the learned vector quantization representation of a video. Specifically, we concatenate the tokens from the multimodal inputs and the target video as a sequence to train the transformer. Tensors obtained from an image and video must be vectorized for concatenation. We do so using the reshape operation (`Reshape`). Therefore, we have a video tensor $\mathbf{z}$ reshaped into a single-index tensor as $\texttt{Reshape}(\mathbf{z}) = [z^{(1)}, \cdots, z^{(hwT)}]$. For simplicity

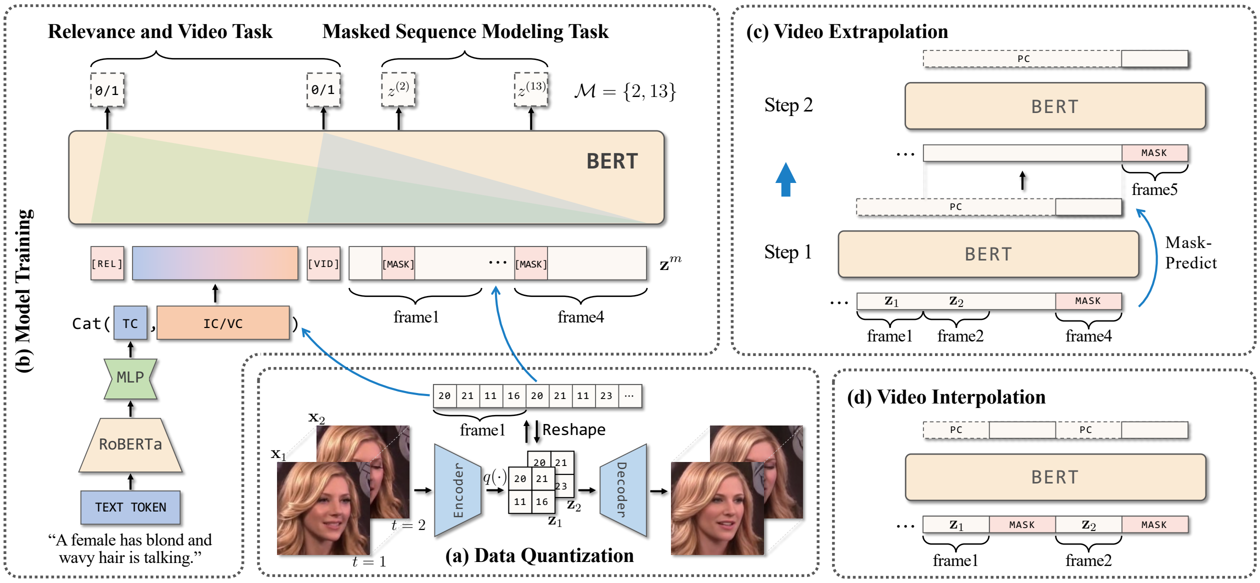


Figure 1. **Pipeline** for training and inference. (a) Data quantization. (b) Model training. Within the BERT module, the green and blue triangles indicate the attention scopes of `[REL]` and `[VID]`, respectively. (c) Video extrapolation. For simplicity, each step represents a full mask-predict process instead of a single forward pass of the transformer. (d) Video interpolation.

of notation, we define $\mathbf{z} \equiv \texttt{Reshape}(\mathbf{z})$. To train the non-autogressive transformer (BERT) on video tokens, we employ three tasks: Masked Sequence Modeling (MSM), RELevance estimation (REL), and VIDeo consistency estimation (VID). During inference, samples are generated via an iterative algorithm based on mask-predict [23], which is simulated by the MSM task during training. The REL and VID tasks regularize the model to synthesize videos that are relevant to the multimodal signals and are temporally consistent. In the following two sections we present each task.

3.1. Masked Sequence Modeling with Relevance

Masked Sequence Modeling. The MSM task is similar to a conditional masked language model [23]. It is essential for the non-autoregressive model to learn bidirectional representations and enables parallel generation (mask-predict). Inspired by M6-UFC [76] and VIMPAC [55], we consider five masking strategies: (I) i.i.d. masking, *i.e.*, randomly masking video tokens according to a Bernoulli distribution; (II) masking all tokens; (III) block masking [55], which masks continuous tokens inside spatio-temporal blocks; (IV) the negation of block masking, which preserves the spatio-temporal block and masks the rest of the tokens; (V) randomly keeping some frames (optional). Strategies I and II are designed to simulate mask-predict sampling (the strategy chosen most of the time). Strategy II helps the model learn to generate from a fully masked sequence in the first step of mask-predict. Strategies III - V can be used as Preservation Control (PC) for preservation tasks, which enable the use of partial images as input (Figs. 3a, 4) and performing long sequence generation (Fig. 5). The MSM task minimizes the softmax cross-entropy loss $\mathcal{L}_{\text{MSM}}$ as follows:

$$\mathcal{L}_{\text{MSM}} = -\frac{1}{|\mathcal{M}|}\sum_{i\in\mathcal{M}} \log P(z^{(i)}|\mathbf{z}^m, \mathbf{c}), \tag{1}$$

where $\mathcal{M}$ is the masking indices, $\mathbf{z}^m$ is the masked sequence, and $\mathbf{c}$ denotes the control sequence.

Relevance Estimation. To encourage the transformer to learn the correlation between multimodal inputs and target videos, we prepend a special token `REL` that is similar to the one used in M6-UFC [76] to the whole sequence, and learn a binary classifier to classify positive and negative sequences. The positive sequence is the same as the sequence used in the MSM task so that we can reuse the same transformer in the forward pass. The negative sequence is constructed by swapping the condition signals along the batch dimension. This swapping does not guarantee constructing strictly negative samples. Nevertheless, we still find it is good enough to make the model learn relevance in practice. The loss function $\mathcal{L}_{\text{REL}}$ for the relevance task is given by:

$$\mathcal{L}_{\text{REL}} = -\log P(1|\mathbf{z}^m, \mathbf{c}) - \log P(0|\mathbf{z}^m, \bar{\mathbf{c}}), \tag{2}$$

where $\bar{\mathbf{c}}$ denotes the swapped control sequence.

3.2. Video Consistency Estimation

To further regularize the model to generate temporally consistent videos, we introduce the video consistency estimation task. Similar to `REL`, we use a special token `VID` to classify positive and negative sequences.

Video Attention. The VID task focuses on video token sequences. Thus, we place the `VID` token between the control and target sequences. We apply a mask to BERT to blind the scope of the `VID` token from the control signals so it only calculates attention from the tokens of the target videos. The positive sequence is the same one used in MSM and REL tasks. The negative sequence is obtained by performing negative augmentation on videos to construct samples that do not have temporally consistent motion or content.

Negative Video Augmentation. We employ four strategies to augment negative video sequences: (I) *frame swapping* – a random frame is replaced by using a frame from another video; (II) *frame shuffling* – frames within a sequence are shuffled; (III) *color jittering* – randomly changing the color of one frame; (IV) *affine transform* – randomly applying an affine transformation on one frame. All augmentations are performed in image space. With $\bar{\mathbf{z}}$ denoting the video sequence after augmentation, the loss $\mathcal{L}_{\text{VID}}$ for the VID task is:

$$\mathcal{L}_{\text{VID}} = -\log P(1|\mathbf{z}^m, \mathbf{c}) - \log P(0|\bar{\mathbf{z}}^m, \mathbf{c}). \qquad (3)$$

Overall, the full objective is $\mathcal{L} = \lambda_{\text{MSM}}\mathcal{L}_{\text{MSM}} + \lambda_{\text{REL}}\mathcal{L}_{\text{REL}} + \lambda_{\text{VID}}\mathcal{L}_{\text{VID}}$, where λs balances the losses.

3.3. Improved Mask-Predict for Video Generation

We employ mask-predict [23] during inference, which iteratively remasks and repredicts low-confidence tokens by starting from a fully-masked sequence. We chose it because it can be used with our bidirectional transformer, as the length of the target sequence is fixed. In addition, mask predict provides several benefits. First, it allows efficient parallel sampling of tokens in a sequence. Second, the unrolling iterations from mask-predict enable direct optimization on synthesized samples, which can reduce exposure bias [46]. Third, information comes from both directions, which makes the generated videos more consistent.

We build our sampling algorithm based on the original mask-predict [23] with two improvements: (I) noise-annealing multinomial sampling, *i.e.*, adding noise during remasking; (II) a new scheme for mask annealing, *i.e.*, using a piecewise linear annealing scheme to prevent the generated motion from being washed out after too many steps of mask-predict. We also apply a beam search from M6-UFC [76]. In Alg. 1, the transformer (`BERT`) takes input tokens $\mathbf{z}_{in}$ and outputs score s and the logits $\tilde{\mathbf{p}}$ for all target tokens. At each mask-predict iteration, we sample tokens with `SampleToken` that returns a predicted token $\mathbf{z}_{out}$ and a vector $\mathbf{y}$ containing its probabilities (unnormalized). `SampleToken` also accepts a scalar σ that indicates the noise level to be added during the token sampling process. `SampleMask`$(\mathbf{y}, \mathbf{m}, N-n)$ remasks n tokens from a total of N tokens according to the multinomial defined by the normalized $\mathbf{y}$, while ensuring tokens with $\mathbf{m} = 1$ are always preserved. $\mathbf{z}_\phi$ denotes the fully-masked sequence. The functions `SampleToken`, `SampleMask` and the schedules of $n^{(i)}$ and $\sigma^{(i)}$ are shown in Appendix.

Algorithm 1 Improved Mask-Predict for Video Generation

Require: Initial PC mask $\mathbf{m}_{\text{PC}}$ and initial token $\mathbf{z}_{in}$.

$\tilde{\mathbf{p}}, s \leftarrow \texttt{BERT}(\mathbf{z}_{in})$
$\mathbf{z}_{out}, \mathbf{y} \leftarrow \texttt{SampleToken}(\tilde{\mathbf{p}}, \sigma^{(1)})$
$\mathbf{z}_{out} \leftarrow \mathbf{m}_{\text{PC}} \odot \mathbf{z}_{in} + (1 - \mathbf{m}_{\text{PC}}) \odot \mathbf{z}_{out}$ ▷ PC
for $i \in \{2, ..., L\}$ **do** ▷ main loop
  for $b \in \{1, ..., B\}$ **do** ▷ beam search
    $\mathbf{m}^b \leftarrow \texttt{SampleMask}(\mathbf{y}, \mathbf{m}_{\text{PC}}, N - n^{(i)})$
    $\mathbf{z}_{in}^b \leftarrow \mathbf{m}^b \odot \mathbf{z}_{out} + (1 - \mathbf{m}^b) \odot \mathbf{z}_\phi$ ▷ remask
    $\tilde{\mathbf{p}}^b, s^b \leftarrow \texttt{BERT}(\mathbf{z}_{in}^b)$ ▷ repredict
  end for
  $b^* \leftarrow \arg\max_b(s^b)$
  $\mathbf{z}_{out}, \mathbf{y} \leftarrow \texttt{SampleToken}(\tilde{\mathbf{p}}^{b^*}, \sigma^{(i)})$
end for
return $\mathbf{z}_{out}$

3.4. Text Augmentation

We study two augmentation methods. First, we randomly drop sentences from the input text to avoid memorizing certain word combinations. Second, we apply a fixed *pretrained* language model, *i.e.*, RoBERTa [33], rather than learning text token embeddings in a lookup table from scratch, to let the model be more robust for input textual information. The features of text tokens are obtained from an additional multilayer perceptron (MLP) appended after the language model that matches the vector dimension with BERT. The features are converted to a weighted sum to get the final embedding of the input text. With the language model, the video generation framework is more robust for out-of-distribution text prompts. When using the tokenizer from an existing work [42], we observed that it might not properly handle synonyms without a common root (Fig. 6).

3.5. Long Sequence Generation

Due to the inherent preservation control mechanism during training (strategy V in the MSM task), we can generate sequences with many more frames than the model is trained with via interpolation or extrapolation. **Interpolation** is conducted by generating the intermediate frames between given frames. As illustrated by Fig. 1 (d), we place $\mathbf{z}_1$ and $\mathbf{z}_2$ at the positions of frames 1 and 3 to serve as preservation controls, *i.e.*, they are kept the same during mask-predict iterations, and we can interpolate a frame between them. **Extrapolation** is similar to interpolation, except we condition the model on previous frames to generate the next frames. As illustrated in Fig. 1 (c), this process can be iterated a number of times to generate longer videos.

4. Experiments

Datasets. We show experiments on the following datasets.

- **Shapes** is proposed by TFGAN [8] for text-to-video generation. Each video shows one object (a geometric shape with specified color and size) displayed in a textured moving background. The motion of an object is described by a text and the background is moving in a random way. There are 30K videos with size 64×64.
- **MUG** [2] contains 52 actors performing 6 different facial expressions. We also provide gender labels for the actors. For a fair comparison, we obtain text descriptions by following TiVGAN [30]. We run experiments on 1039 videos with resolution 128×128.
- **iPER** [32] consists of 206 videos of 30 subjects wearing different clothes performing an A-pose and random actions. Experiments are conducted with size 128×128.
- **Multimodal VoxCeleb** is a new dataset for multimodal video generation. We first obtain $19,522$ videos from VoxCeleb [37] after performing pre-processing [53]. Second, we manually label 36 facial attributes described in CelebA [34] for each video. Third, we use a probabilistic context-free grammar to generate language descriptions [71]. Finally, we run APDrawingGAN [73] to get artistic portrait drawings and utilize face-parsing [74] to produce segmentation masks.

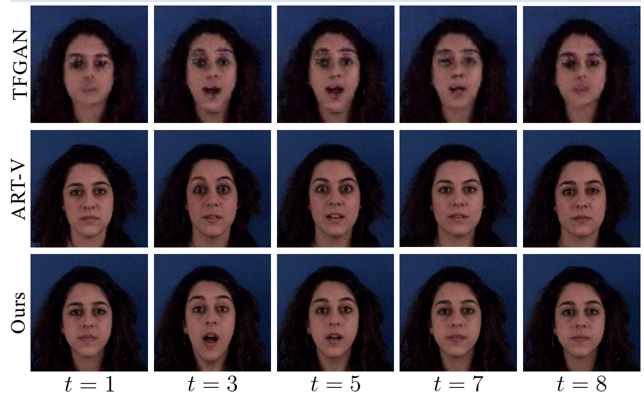
Baseline Methods. We run TFGAN [8] on Shapes, MUG, and Multimodal VoxCeleb datases for comparison of text-to-video synthesis. We also compare our approach with TiVGAN [30] on MUG. Additionally, we unify the autoregressive transformer of DALL-E [45] and the autoencoder from VQGAN (the same one used in our method) in a multimodal video generative model. We name the strong baseline as **A**uto**R**egressive **T**ransformer for **V**ideo generation (**ART-V**) and compare it with our bidirectional transformer for predicting video tokens. We train ART-V with the next-token-prediction objective on concatenated token sequences obtained from input controls and target videos.

Evaluation Metrics. We follow the metrics from existing works on Shapes and MUG to get a fair comparison. Specifically, we compute classification accuracy on Shapes and MUG and Inception Score (IS) [52] on MUG. On Multimodal VoxCeleb and iPER datasets, we report Fréchet Video Distances (FVD) [58] that is computed from 2048 samples, and Precision-Recall Distribution (PRD) (F_8 and $F_{1/8}$) for diversity [51]. We further perform human evaluation using Amazon Mechanical Turk to verify the quality and diversity of synthesized videos.

4.1. Text-to-Video Generation

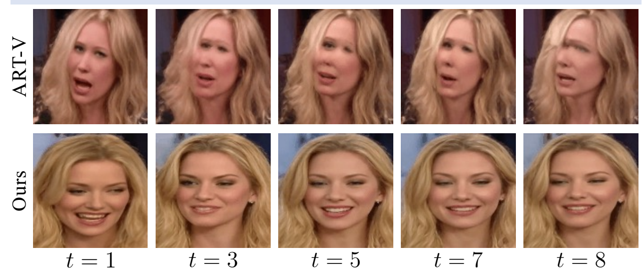
Shapes. We report the classification accuracy in Tab. 1 (top four rows) for the Shapes dataset. ART-V and MMVID are

A woman is making a surprise face.



(a) Samples on **MUG**. ART-V and our method can generate sharp and temporally consistent videos while frames produced by TFGAN are blurry.

A female is wearing lipstick. This person has wavy hair and blond hair. She has high cheekbones, arched eyebrows and rosy cheeks. She has heavy makeup. She is young.



(b) Samples on **Multimodal VoxCeleb**. Frame generated by ART-V at $t=1$ is sharp and clear, but are blurry at later steps such as $t=5$ and 8.

Figure 2. **Text-to-video** generation results for different methods. Sample frames are shown at several time steps (t). Conditioned text is provided at the top with light blue background.

Table 1. Classification accuracy (%) on the Shapes dataset for video generation. Our method achieves the best performance.

Condition	Methods	Shape	Color	Size	Motion	Dir	Avg
Text Only	TFGAN [8]	80.22	**100.00**	84.33	**99.90**	**99.95**	92.88
	ART-V	95.07	98.68	97.71	92.72	96.04	96.04
	MMVID (Ours)	**95.56**	99.71	**97.95**	97.80	99.61	**98.12**
Multimodal	ART-V	92.82	97.17	97.31	89.55	93.99	94.17
	MMVID (Ours)	**98.19**	**99.76**	**98.83**	**99.46**	**99.95**	**99.24**

trained for 100K iterations. Compared with TFGAN [8], our model achieves significantly higher classification accuracy for Shape, Size, and Average (Avg) categories. Compared with ART-V, we perform better in all the categories. Note that our method has slightly lower accuracy on Color, Motion, and Direction (Dir) than TFGAN. The reason might be the VQGAN introduces errors in reconstruction, since the background is diverse and moving, and it needs to encode these small translations by only 4×4 codes. Note that to have a fair comparison, we do not apply text augmentation when performing comparison with other works.

MUG. We follow the experimental setup in TiVGAN [30]

Table 2. Inception Score (IS) and classification accuracy (%) on MUG for video generation. We mark '*' to IS values reported in TiVGAN. Our model achieves highest accuracy and IS.

Condition	Methods	Gender (%) ↑	Expression (%) ↑	IS ↑
Text Only	TGAN [49]	-	-	*4.63
	MoCoGAN [57]	-	-	*4.92
	TGANs-C [40]	-	-	*4.65
	TiVGAN [30]	-	-	*5.34
	TFGAN [8]	99.22	**100.00**	5.53
	ART-V	93.46	99.12	5.72
	MMVID (Ours)	**99.90**	**100.00**	**5.94**
Multimodal	ART-V	89.16	98.54	5.59
	MMVID (Ours)	**98.14**	**100.00**	**5.85**

Table 3. Evaluation metrics for text-to-video generation on iPER and Multimodal VoxCeleb datasets.

Dataset	Methods	Human (%) ↑	FVD ↓	F_8 ↑	$F_{1/8}$ ↑
iPER	ART-V	-	277.604	0.936	0.806
	MMVID (Ours)	-	**209.127**	**0.944**	**0.924**
Multimodal VoxCeleb	ART-V	46.0	70.952	0.946	0.940
	MMVID (Ours)	**54.0**	**59.464**	**0.965**	**0.961**

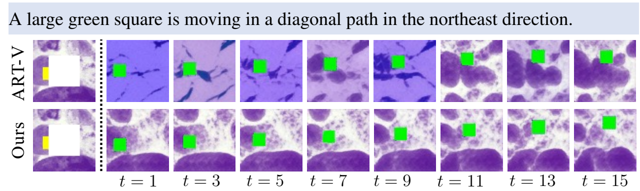
for experiments on the MUG expression dataset. We train models with a temporal step size of 8 due to the memory limit of GPU. Note TiVGAN is trained with a step size of 4 and generates 16-frame videos, while our model generates 8-frame videos in a single forward. We also train a 3D ConvNet as described in TiVGAN to evaluate the Inception Score and perform classification on Gender and Expression. Results are shown in Fig. 2a and Tab. 2 (top 8 rows). Our model achieves the best performance.

iPER. We show the results of the dataset in Tab. 3 (top 3 rows), demonstrating the advantages of our method. Long sequence generation results are shown in Fig. 5.

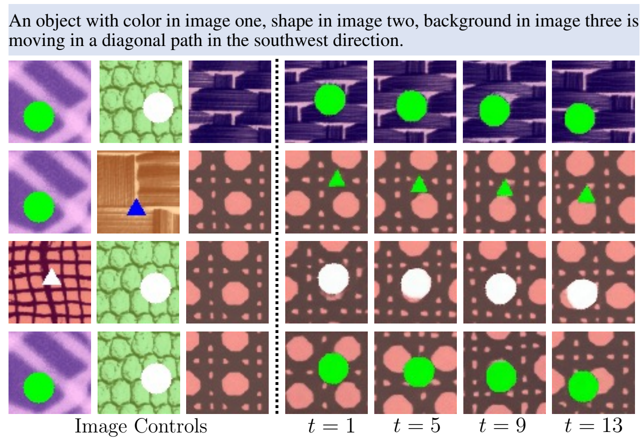
Multimodal VoxCeleb. We train ART-V and our model at a spatial resolution of 128×128 and a temporal step of 4 to generate 8 frames. Our method shows better results than ART-V on all the metrics, as shown in Tab. 3 (bottom two rows). We notice ART-V can also generate video samples with good visual quality and are aligned well with the text descriptions. However, ART-V often produces samples that are not temporally consistent. For example, as shown in Fig. 2b, the frame generated by ART-V at $t = 1$ is sharp and clear, but frames at $t = 5$ or $t = 8$ are blurry. The reason might be the exposure bias in autoregressive models becomes more obvious as the sequence length is long, *i.e.*, an 8-frame video at resolution 128×128 has 512 tokens. Thanks to bidirectional information during training and inference, our MMVID is able to produce temporally consistent videos.

4.2. Multimodal Video Generation

Multimodal conditions can evolved in two cases: independent and dependent, and we show experiments on both.



(a) **Independent** multimodal control. The text description specifies the size, color, and shape of the object, and its motion. The visual control is a *partially* observed image with its center masked out (shown as white), which provides control for the background. ART-V can generate correct object and motion, but suffers from incorporating consistent visual inputs such that the background is not temporal consistent.



(b) **Dependent** multimodal controls. The text description specifies from which image to extract color, shape, and background.

Figure 3. **Multimodal** generation results on Shapes with *textual* (at top) and *visual* (first column(s)) modalities. Sample frames are shown at several time step (t).

Independent Multimodal Controls. This setting is similar to conventional conditional video generation, except the condition is changed to multimodal controls. We conduct experiments on Shapes and MUG datasets with the input condition as the combination of text and image. The bottom two rows in Tab. 1 and Tab. 2 demonstrate the advantages of our method over ART-V on all metrics. Additionally, we provide generated samples in Fig. 3a, where only a *partial* image is given as the visual condition. As can be seen, ART-V cannot satisfy the visual constraint well and the generated video is not consistent. The quality degradation for multimodal video synthesis of ART-V can also be verified in Tab. 1 as it shows lower classification accuracy than text-only generation, while our method is able to generate high quality videos for different condition signals.

We also conduct extensive experiments of video generation under various combinations of textual and image controls on Multimodal VoxCeleb, as shown in Fig. 4. We apply three different image controls, including segmentation mask (Fig. 4 row (a) - (b)), drawing (Fig. 4 row (c) - (d)), and partial image (Fig. 4 row (d) - (f)). In Fig. 4 row (b), our

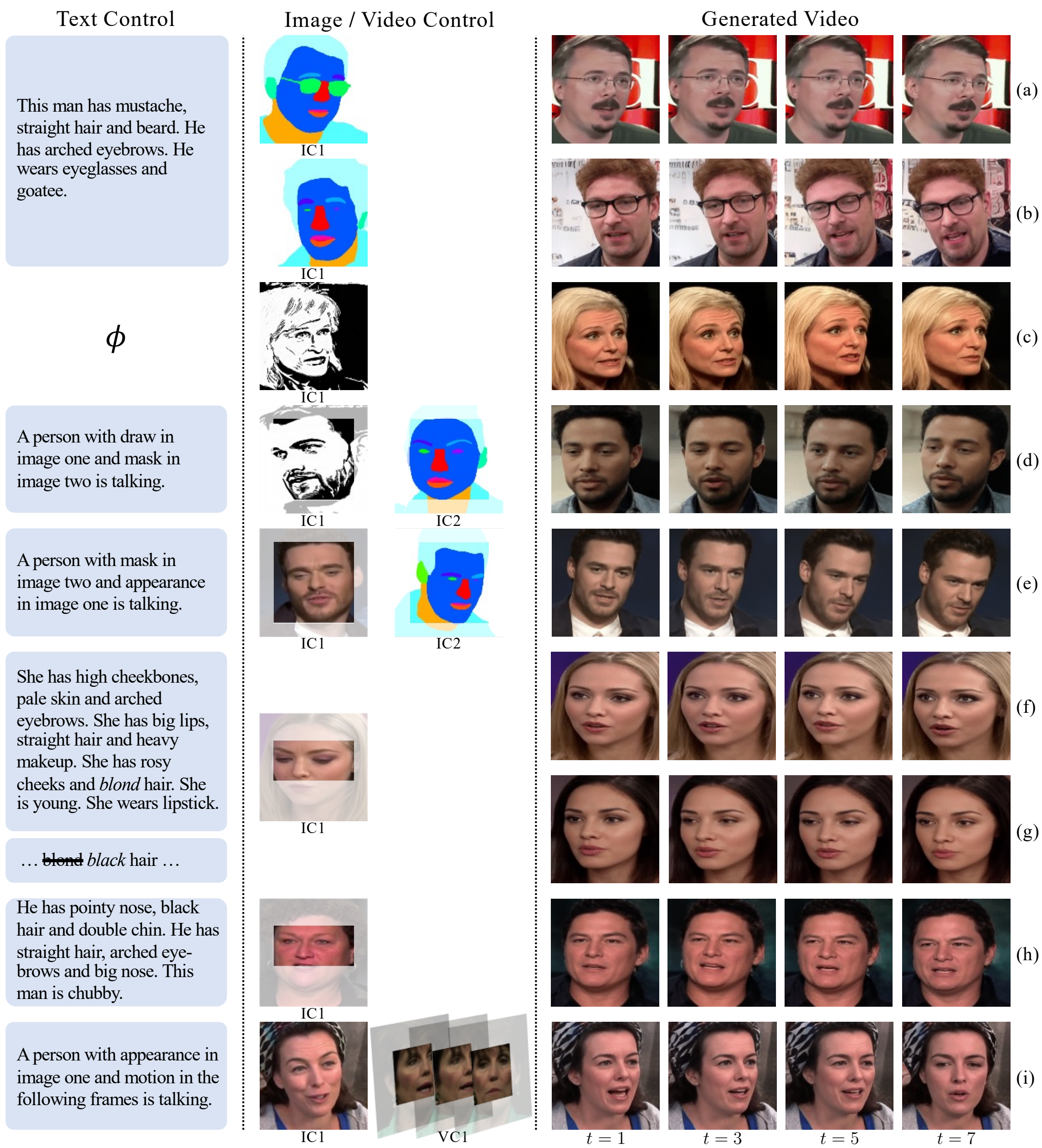


Figure 4. **Independent and Dependent** multimodel video generation on Multimodal VoxCeleb with textual control (TC), image control (IC), and video control (VC). *Row (a) - (b)*: TC + IC (segmentation mask); *Row (c)*: TC (null) + IC (drawing); *Row (d) - (e)*: dependent TC + IC; *Row (f) - (h)*: TC + IC (partial image) and the TC of (g) is obtained from the TC of (f) by replacing "blond" with "black"; *Row (i)*: dependent TC + VC and the VC includes content and motion information.

method can synthesize frames with eyeglasses even though eyeglasses are not shown in segmentation mask. In Fig. 4 row (f) - (g), we show that using the same image control while replacing the "blond" with "black" in the text description, we can generate frames with similar content except the hair color is changed. Such examples demonstrate that our method has a good understanding of multimodal controls.

Dependent Multimodal Controls. Furthermore, we introduce a novel task for multimodal video generation where textual and visual controls are dependent, such that the actual control signals are guided by the textual description. For example, Fig. 3b illustrates how the text informs from which image the model should query color, shape, and background information. More synthesized examples on Multimodal VoxCeleb are given in Fig. 4. For Fig. 4 row (d) - (e), our model learns to combine detailed facial features

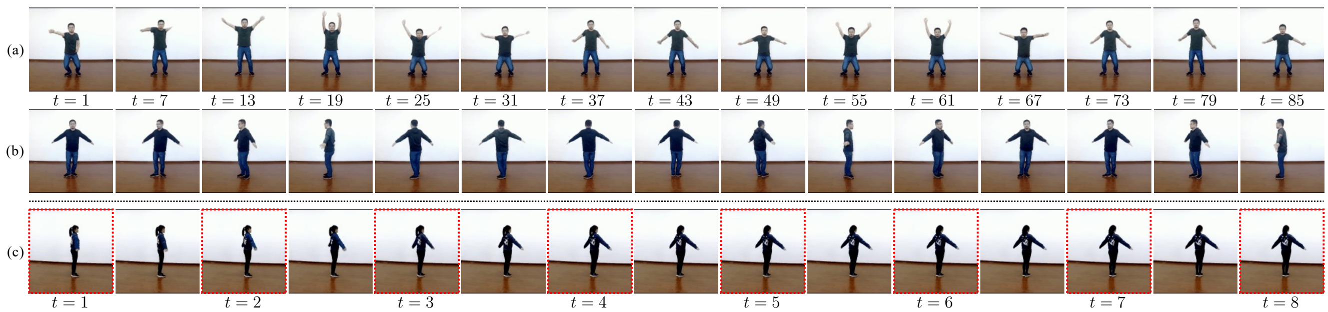


Figure 5. **Extrapolation and Interpolation**. *Row (a) - (b)*: long sequence generation via extrapolation. *Row (c)*: interpolating a real sequence. Frames in dotted red boxes are fixed as preservation control. Textual controls for each row are: (a) "Person 024 dressed in 2 is performing random pose, normal speed."; (b) "Person 024 dressed in 1 is performing A-pose, normal speed."; and (c) "Person 028 dressed in 2 is performing A-pose, normal speed."

from drawing or image and coarse features (*i.e.*, pose) from mask. For Fig. 4 row (i), our method successfully retargets the subject with an appearance from the given image control (IC1) and generates frames with the motion specified by consecutive images that provide motion control (VC1).

Table 4. Analysis on Shapes for video augmentation strategies.

Video Augmentation				Accuracy (%)					
Swap	Shuffle	Color	Affine	Shape	Color	Size	Motion	Dir	Avg
✗	✗	✗	✗	90.43	89.07	95.61	92.48	99.13	93.34
✓	✗	✗	✗	91.02	89.84	93.75	91.02	98.05	92.73
✗	✓	✗	✗	88.28	89.45	94.53	88.28	98.44	91.80
✗	✗	✓	✗	91.80	**91.02**	94.53	93.36	98.83	93.91
✗	✗	✗	✓	90.62	90.62	95.31	89.84	98.83	93.05
✓	✓	✓	✓	**93.36**	88.28	**95.70**	**93.75**	**99.61**	**94.14**

4.3. Long Sequence Generation and Ablation

Long Sequence Generation. Our approach enables temporal extrapolation of videos. We show samples of video extrapolation and interpolation in Fig. 5. Samples from Fig. 5 row (a) - (b) are generated by being iteratively conditioned on previous 6 frames to generate the following 2 frames. Fig. 5 row (c) shows an example of synthesizing one frame by interpolating two consecutive real frames.

Analysis on VID Task. We perform analysis for different `VID` strategies on the Shapes dataset. Tab. 4 shows that the highest average accuracy is achieved when all augmentation is used (sampled uniformly). Also note that accuracy for color is the highest when we only apply color augmentation.

Analysis on Language Embedding. Analysis of using a pretrained language model is shown in Fig. 6. The method with a language model (*w/* RoBERTa) is more robust to various text inputs than the one without it (*w/o* RoBERTa).

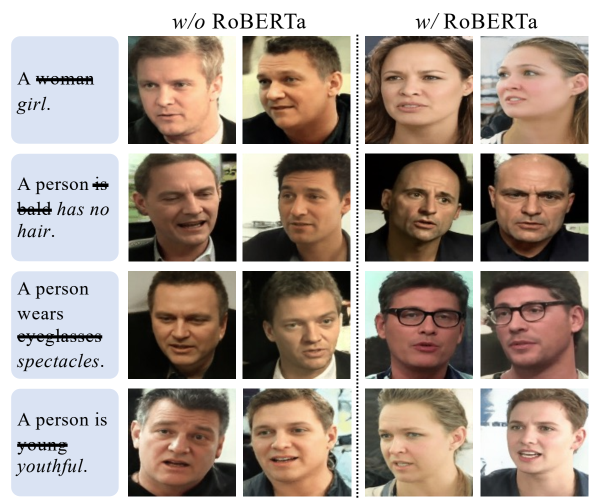


Figure 6. **Analysis on language embedding**. Samples are generated with out-of-distribution textual inputs. We reword the original text (strikethrough) with equivalent descriptions (italic) that do not exist in the training. We show the first frames from the generated sequences for each method. Frames generated using the pretrained language model (*w/* RoBERTa) is more correlated with text inputs.

5. Limitation and Conclusion

This paper targets a new problem, which is video generation using multimodal inputs. To tackle the problem, we utilize a two-stage video generation framework that includes an autoencoder for quantized representation of images and videos and a non-autoregressive transformer for predicting video tokens from multimodal input signals. Several techniques are proposed, including the special `VID` token, textual embedding, and improved mask prediction, to help generate temporally consistent videos. On the other hand, the proposed method also contains some limitations, including temporal consistency issues for high-resolution videos, generating diverse motion patterns for longer sequences, and further improving the diversity of non-autoregressive transformers. More details can be found in the Appendix. Besides improving the limitation, a future direction might be to leverage more control modalities, such as audio, to generate videos with a much higher resolution.

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. *arXiv:1810.02419*, 2018. 2

[2] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4. IEEE, 2010. 5

[3] Adil Kaan Akan, Erkut Erdem, Aykut Erdem, and Fatma Güney. Slamp: Stochastic latent appearance and motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14728–14737, 2021. 2

[4] Adil Kaan Akan, Sadra Safadoust, Erkut Erdem, Aykut Erdem, and Fatma Güney. Stochastic video prediction with structure and motion. *arXiv preprint arXiv:2203.10528*, 2022. 2

[5] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *ICLR*, 2017. 2

[6] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021. 2

[7] 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. 1, 2

[8] 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 *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1995–2001. International Joint Conferences on Artificial Intelligence Organization, 2019. 1, 5, 6

[9] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Bjorn Ommer. Understanding object dynamics for interactive image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2

[11] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *ECCV*, 2018. 2

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*. PMLR, 2020. 1, 2

[13] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *arXiv preprint arXiv:2009.11278*, 2020. 2

[14] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv*, 2019. 2

[15] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018. 2

[16] Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, 2017. 2

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[18] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*, 2021. 2

[19] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Bjorn Ommer. Stochastic image-to-video synthesis using cinns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[21] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. 1, 2

[22] Junlong Gao, Xi Meng, Shiqi Wang, Xia Li, Shanshe Wang, Siwei Ma, and Wen Gao. Masked non-autoregressive image captioning. *arXiv preprint arXiv:1906.00717*, 2019. 2

[23] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019. 2, 3, 4

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[25] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017. 2

[26] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. In *NeurIPS*, 2018. 2

[27] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[28] Sangeek Hyun, Jihwan Kim, and Jae-Pil Heo. Self-supervised video gans: Learning for appearance consistency and motion coherency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10826–10835, 2021. 2

[29] Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *Neural Networks*, 132:506–520, 2020. 2

[30] Doyeon Kim, Donggyu Joo, and Junmo Kim. Tivgan: Text to image to video generation with step-by-step evolutionary generator. *IEEE Access*, 8:153113–153122, 2020. 1, 2, 5, 6

[31] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, 2018. 2

[32] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 5

[33] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[35] Elman Mansimov, Alex Wang, Sean Welleck, and Kyunghyun Cho. A generalized framework of sequence generation with application to undirected sequence models. *arXiv preprint arXiv:1905.12790*, 2019. 2

[36] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubramanian. Attentive semantic video generation using captions. In *ICCV*, 2017. 2

[37] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 2020. 2, 5

[38] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017. 1, 2

[39] Junting Pan, Chengyu Wang, Xu Jia, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang. Video generation from single semantic label map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2019. 2

[40] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, 2017. 2, 6

[41] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4

[43] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[44] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. *arXiv preprint arXiv:2006.10704*, 2020. 2

[45] 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. 2, 5

[46] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 4

[47] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 1, 2

[48] Jian Ren, Menglei Chai, Oliver J Woodford, Kyle Olszewski, and Sergey Tulyakov. Flow guided transformable bottleneck networks for motion retargeting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10795–10805, 2021. 2

[49] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *ICCV*, 2017. 2, 6

[50] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. *IJCV*, 2020. 2

[51] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lučić, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 5

[53] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 5

[54] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021. 2

[55] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021. 3

[56] 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 on Learning Representations*, 2021. 1, 2

[57] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, 2018. 1, 2, 6

[58] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. To-

wards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 5

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017. 2

[60] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *ICLR*, 2017. 2

[61] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017. 2

[62] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. In *NeurIPS*, 2016. 1, 2

[63] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017. 2

[64] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021. 2

[65] Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019. 2

[66] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5264–5273, 2020. 2

[67] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020. 2

[68] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *ICLR*, 2020. 2

[69] 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. 2

[70] Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. Generative video transformer: Can objects be the words? In *ICML*, 2021. 1, 2

[71] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[72] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[73] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '20)*, pages 8214–8222, 2020. 5

[74] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, pages 1–18, 2021. 5

[75] Vladyslav Yushchenko, Nikita Araslanov, and Stefan Roth. Markov decision process for video generation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[76] Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. Ufcbert: Unifying multi-modal controls for conditional image synthesis. *arXiv preprint arXiv:2105.14211*, 2021. 1, 2, 3, 4