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

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

[Motivation]

Multimodal conditional video synthesis

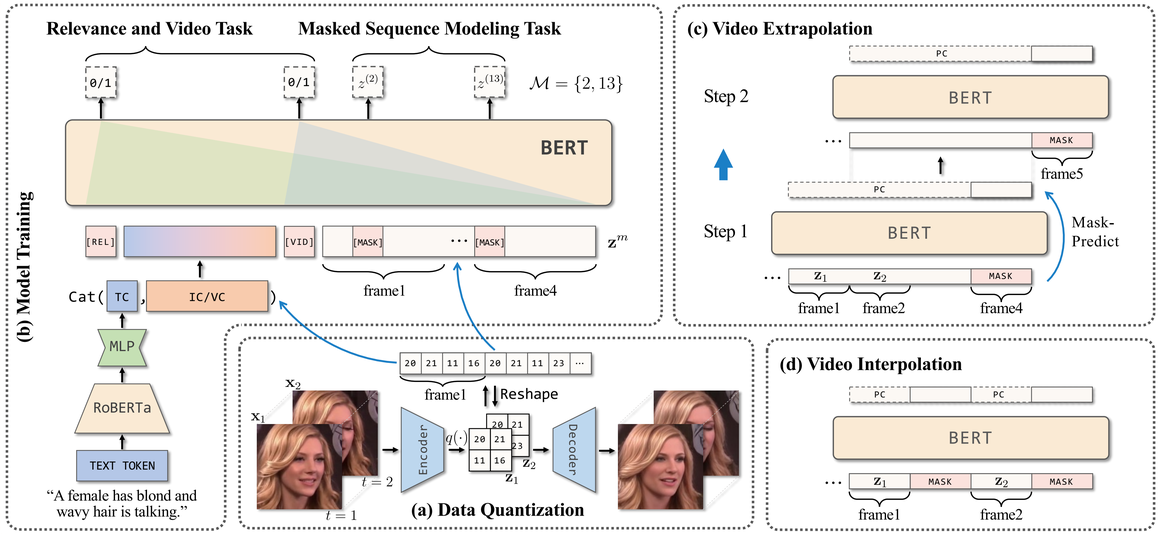
- Most methods for conditional video synthesis use a ***single*** modality :
  - ➢ it is problematic to condition on an image but to generate a specific motion trajectory desired by the user.
  - ➢ language information can describe the desired motion but not precisely defining the content of the video.
- This work presents a **M**ulti**M**odal **VID**eo generation framework (**MMVID**) that benefits from text and images provided jointly or separately.

[Method]

Two-Stage Video Generation

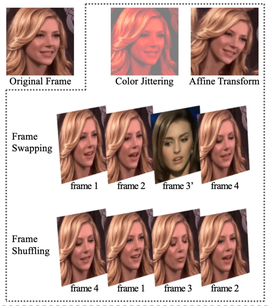
- **First**: an AutoEncoder (VQ-GAN / VQ-VAE) to obtain a *quantized* representation for images.
- **Second** : a *bidirectional* transformer for modeling the correlation between multimodal controls and the learned vector quantization representation of a video.
  - ➢ Masked Sequence Modeling (**MSM**): on target video sequence.
  - ➢ Relevance Estimation (**REL**).
  - ➢ Video Consistency Estimation (**VID**).
  - ➢ Improved **Mask-Predict** for inference.
  - ➢ **Text Augmentation**.

[Method] Pipeline for Training and Inference

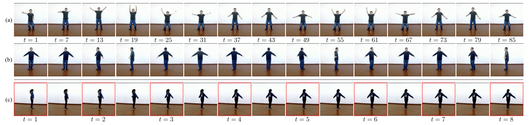


VID token

- Video Attention: apply a mask to BERT to blind the scope of the VID token from the control signals
- Negative Video Augmentation



iPER Dataset



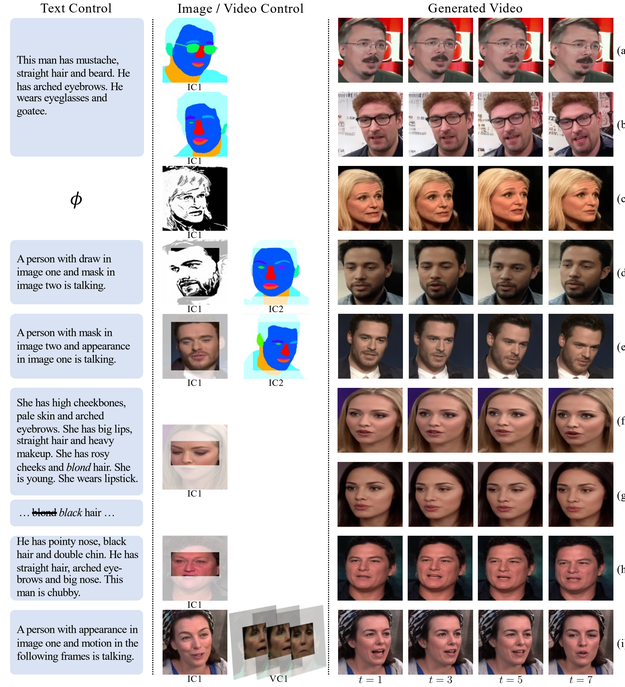
Mask-Predict

Algorithm 1 Improved Mask-Predict for Video Generation

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

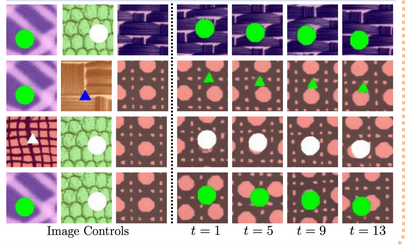
$\tilde{\mathbf{p}}, s \leftarrow \texttt{BERT}(\mathbf{z}_{in})$
$\mathbf{z}_{out}, \mathbf{y} \leftarrow \texttt{SampleToken}(\tilde{\mathbf{p}}, \sigma^{(1)})$
$\mathbf{z}_{out} \leftarrow \mathbf{m}_{\text{PC}} \odot \mathbf{z}_{in} + (1 - \mathbf{m}_{\text{PC}}) \odot \mathbf{z}_{out}$ ▷ PC
for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

[Results] Multimodal VoxCeleb Dataset



Shapes Dataset

An object with color in image one, shape in image two, background in image three is moving in a diagonal path in the southwest direction.



Project Webpage