

# Deep Video Generation, Prediction and Completion of Human Action Sequences

Haoye Cai<sup>1,3</sup>[0000-0001-7041-563X], Chunyan Bai<sup>1,4</sup>[0000-0001-5431-795X],  
Yu-Wing Tai<sup>2</sup>[0000-0002-3148-0380], and Chi-Keung Tang<sup>1</sup>[0000-0001-6495-3685]

<sup>1</sup> Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong  
{hcai aa, cbai }@connect.ust.hk cktang@cs.ust.hk

<sup>2</sup> Tencent Youtu, Shenzhen, China  
yuwingtai@tencent.com

<sup>3</sup> Stanford University, Stanford, CA 94305, USA

<sup>4</sup> Carnegie Mellon University, Pittsburgh, PA 15213, USA  
Project page: <https://iamacewhitel.github.io/supp>

**Abstract.** Current video generation/prediction/completion results are limited, due to the severe ill-posedness inherent in these three problems. In this paper, we focus on human action videos, and propose a general, two-stage deep framework to generate human action videos with no constraints or arbitrary number of constraints, which uniformly addresses the three problems: video generation given no input frames, video prediction given the first few frames, and video completion given the first and last frames. To solve video generation from scratch, we build a two-stage framework where we first train a deep generative model that generates human pose sequences from random noise, and then train a skeleton-to-image network to synthesize human action videos given the human pose sequences generated. To solve video prediction and completion, we exploit our trained model and conduct optimization over the latent space to generate videos that best suit the given input frame constraints. With our novel method, we sidestep the original ill-posed problems and produce for the first time high-quality video generation/prediction/completion results of much longer duration. We present quantitative and qualitative evaluations to show that our approach outperforms state-of-the-art methods in all three tasks.

**Keywords:** Video Generation · Generative Models

## 1 Introduction

In this paper we propose a general, two-stage deep framework for human video generation (i.e. generating video clips directly from latent vectors), prediction (i.e. predicting future frames of a short clip or single frame), and completion (i.e. completing the intermediate content given the beginning and the ending),

---

Equal Contribution.



































