RECENT ADVANCES IN OPEN-SOURCE VIDEO GENERATION AND EDITING

Qifeng Chen
The Hong Kong University of Science and Technology
Sora by OpenAI
Sora + Our Seeing-and-Hearing
Open-source Video Generation and Editing

**Video Foundation Models**

- Single Domain
- Open Domain
- World Simulator

**Advanced Functions**

- Controllability
- Multimodality
- Editing

- Character
- Local Region
- Pose
- Visual + Audio
- Temporal Consistency

- Video Generation with High Dynamics
- Multimodality Joint Generation
VideoCrafter
An Open-source Toolkit for Text-to-Video Generation and Editing
VideoCrafter1: Open Diffusion Models for High-Quality Video Generation

Haoxin Chen 1, Menghan Xia 1, Yingqing He 1,2, Yong Zhang 1, Xiaodong Cun 1
Shaoshu Yang 1,3, Jinbo Xing 1,4, Yaofang Liu 1,5, Qifeng Chen 2,
Xintao Wang 1, Chao Weng 1, Ying Shan 1

1 Tencent AI Lab 2 Hong Kong University of Science and Technology
3 Chinese Academy of Sciences 4 The Chinese University of Hong Kong 5 City University of Hong Kong

Diffusion Models for Video Generation

- Make-A-Video @Meta
- Imagen Video @Google
- Gen1 @Runway
- ModelScope @Alibaba
- VideoComposer @Alibaba
- Gen2 @Runway
- Pikalab @Pikalab
- Genmo @Genmo.ai
- LaVie @GH AI Lab
- Show-1 @NUS

- VDM @Google
- MagicVideo @ByteDance
- LVDM @Tencent
- VideoCrafter0.9 @Tencent
- Video LDM @NVIDIA
- MorphStudio @MorphStudio
- i2VGen-XL @Alibaba
- MoonValley @MoonValley

2022/09-10 | 2023/02-03 | 2023/06 | 2023/09 | 2023/10

Open Source | Available Online
## Evaluation Quality Dimension

- subject consistency
- background consistency
- motion smoothness
- aesthetic quality
- imaging quality
- dynamic degree

### Model Name (clickable) vs Quality Score vs Selected Score vs subject consistency vs background consistency

<table>
<thead>
<tr>
<th>Model Name (clickable)</th>
<th>Quality Score</th>
<th>Selected Score</th>
<th>subject consistency</th>
<th>background consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sora</td>
<td>79.69%</td>
<td>79.69%</td>
<td>94.96%</td>
<td>95.84%</td>
</tr>
<tr>
<td>Gen-2</td>
<td>78.79%</td>
<td>78.79%</td>
<td>97.61%</td>
<td>97.61%</td>
</tr>
<tr>
<td>Pika</td>
<td>78.26%</td>
<td>78.26%</td>
<td>96.76%</td>
<td>98.95%</td>
</tr>
<tr>
<td>VideoCrafter-1.0</td>
<td>78.14%</td>
<td>78.14%</td>
<td>95.1%</td>
<td>98.04%</td>
</tr>
<tr>
<td>Show-1</td>
<td>77.5%</td>
<td>77.5%</td>
<td>95.53%</td>
<td>98.02%</td>
</tr>
<tr>
<td>LaVie-Interpolation</td>
<td>75.86%</td>
<td>75.86%</td>
<td>92.6%</td>
<td>97.33%</td>
</tr>
<tr>
<td>LaVie</td>
<td>75.75%</td>
<td>75.75%</td>
<td>91.41%</td>
<td>97.47%</td>
</tr>
<tr>
<td>ModelScope</td>
<td>74.91%</td>
<td>74.91%</td>
<td>89.87%</td>
<td>95.29%</td>
</tr>
<tr>
<td>VideoCrafter-0.9</td>
<td>71.53%</td>
<td>71.53%</td>
<td>86.24%</td>
<td>92.88%</td>
</tr>
</tbody>
</table>
Foundation Models in Video Generation: VideoCrafter1

VideoCrafter1
TEXT TO VIDEO
Foundation Models in Video Generation: VideoCrafter1

Latent Video Diffusion Models (LVDM 2022.11)
• 1024 × 576 resolution
• Image Video Joint Training: 20 million videos + 600 million images.
• Text-to-Video + Image-to-Video

• The first open-source generic I2V model that can strictly preserve the content and structure of the input reference image while animating it into a video.

Latent Video Diffusion Models for High-Fidelity Long Video Generation

Yingqing He1, Tianyu Yang2, Yong Zhang2, Ying Shan2, Qiheng Chen1

1The Hong Kong University of Science and Technology 2Tencent AI Lab
ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models

Yingqing He\textsuperscript{*1} Shaoshu Yang\textsuperscript{*2} Haoxin Chen\textsuperscript{3} Xiaodong Cun\textsuperscript{3} Menghan Xia\textsuperscript{3} Yong Zhang\textsuperscript{*3} Xintao Wang\textsuperscript{3} Ran He\textsuperscript{2} Qifeng Chen\textsuperscript{*1} Ying Shan\textsuperscript{3}

\textsuperscript{1} Hong Kong University of Science and Technology \textsuperscript{2} Chinese Academy of Sciences \textsuperscript{3} Tencent AI Lab

(\textsuperscript{*} Equal Contribution \textsuperscript{#} Corresponding Author)

ICLR 2024 Spotlight

4K/2K Image/Video Generation
Text-to-2K videos (2048 × 1152)
Text-to-2K videos (2048 × 1152)
Text-to-2K videos (2048 × 1152)
Text-to-2K videos (2048 × 1152)
Text-to-2K videos (2048 × 1152)
Text-to-2K videos (2048 × 1152)
Text-to-4K Images
Controllability
High-Controllability: Animate-A-Story for Character Control
Demo

Controllable Character

A day of a Teddy bear

A teddy bear wakes up in the morning in his bed.

Duck Kingdom

There is a Duck Kingdom
High-Controllability: Follow-Your-Click (in submission)
High-Controllability: Follow-Your-Click
FOLLOW-YOUR-POSE: POSE-GUIDED TEXT-TO-VIDEO GENERATION USING POSE-FREE VIDEOS

AAAI 2024
High-Controllability: Follow-Your-Pose
CODEF: CONTENT DEFORMATION FIELDS FOR TEMPORALLY CONSISTENT VIDEO PROCESSING
CVPR 2024 (HIGHLIGHT)

Hao Ouyang*, Qiuyu Wang*, Yuxi Xiao*, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen†, Yujun Shen† (*equal contribution, †corresponding author)
FATEZERO: FUSING ATTENTIONS FOR ZERO-SHOT TEXT-BASED VIDEO EDITING
ICCV 2023 (ORAL)

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen
Results

"silver jeep ➔ posche car"
Shape Editing

"+ Van Gogh style"
Style Editing

"squirrel, Carrot ➔ rabbit, eggplant"
Attribute Editing
Thank you!
More information at https://cqf.io