

Show-1: Marrying Pixel and Latent Diffusion Models for Text-to-Video Generation

Qin Zhao
2023.11.20

Outlets

- Introduction
- Method
- Experiment
- Demo results

Introduction

Significant advancements have been achieved in the realm of large-scale pretrained text-to-video Diffusion Models (VDMs).

These VDMs can be classified into two types:

- (1) **Pixel-based VDMs** that directly denoise pixel values, including Make-A-Video, Imagen Video, PYoCo.
- (2) **Latent-based VDMs** that manipulate the compacted latent space within a variational autoencoder (VAE), like Video LDM and MagicVideo .

Introduction

A blue tiger in the grass in the sunset, surrounded by butterflies.



Pixel-based VDM
64 x 40

Latent-based VDM
64 x 40

Latent-based VDM
256 x 160

“A wolf drinking coffee in a café.”



Pixel-based VDM
64 x 40

Latent-based VDM
64 x 40

Latent-based VDM
256 x 160

However, both of these two methods have pros and cons.

Pixel-based VDMs:

Motion accurately aligned but demand **expensive computational costs**.

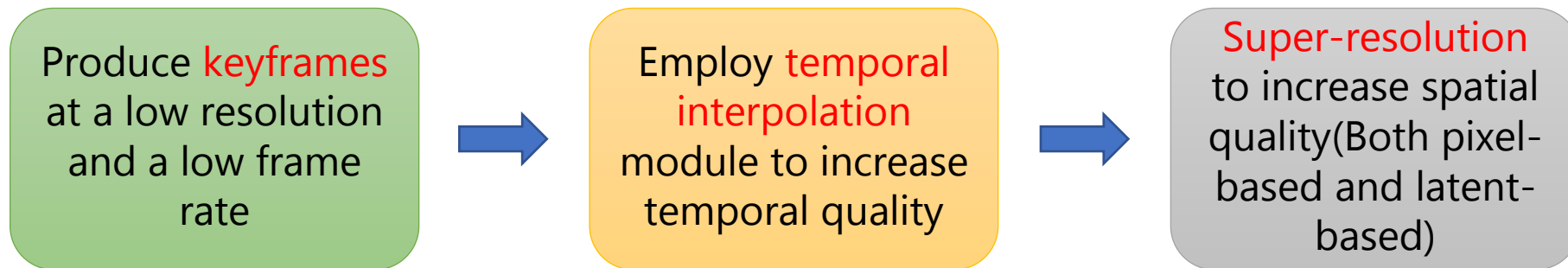
Latent-based VDMs:

More resource-efficient (Reduced-dimension latent space), but can't work well in **small latent space**. Furthermore, this latent model will focus more on spatial appearance but may also **ignore the text-video alignment**.

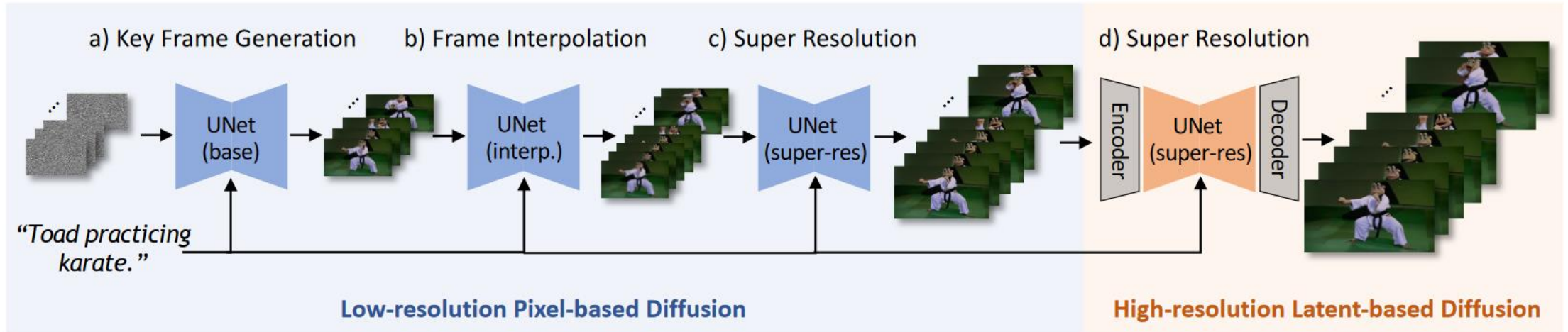
Introduction

To marry the strength and alleviate the weakness of pixel-based and latent-based VDMs, this paper put forward Show-1, which is an efficient text-to-video model that generates videos of not only decent video-text alignment but also high visual quality.

Show-1 follows the conventional coarse-to-fine video generation pipeline.



Method



Employs pixel-based VDMs for the keyframe module and the temporal interpolation module at a low resolution, producing key frames of precise text-video alignment and natural motion with low computational cost.

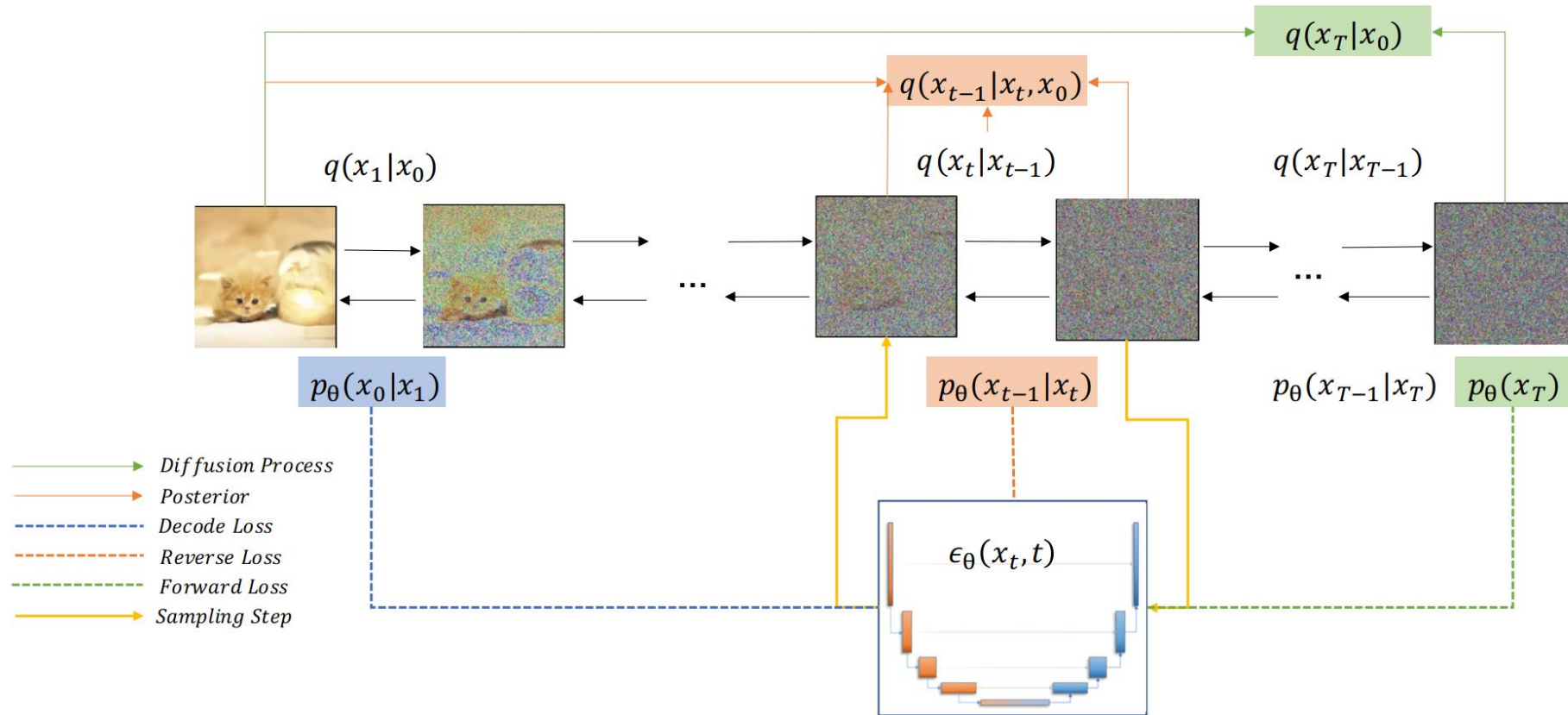
With respect to super-resolution module, Show-1 proposes a novel two-stage super-resolution module. Firstly, employs pixel-based VDMs to upsample the video from 64×40 to 256×160 . When it comes to the second stage, a novel expert translation module based on latent-based VDMs is designed to further upsample it to 572×320 with low computation cost.

To sum up, Pixel-based VDMs produce videos of lower resolution with better text-video alignment, while latent-based VDMs upscale these low-resolution videos from Pixel-based VDMs to then create high-resolution videos with low computation cost.

Method

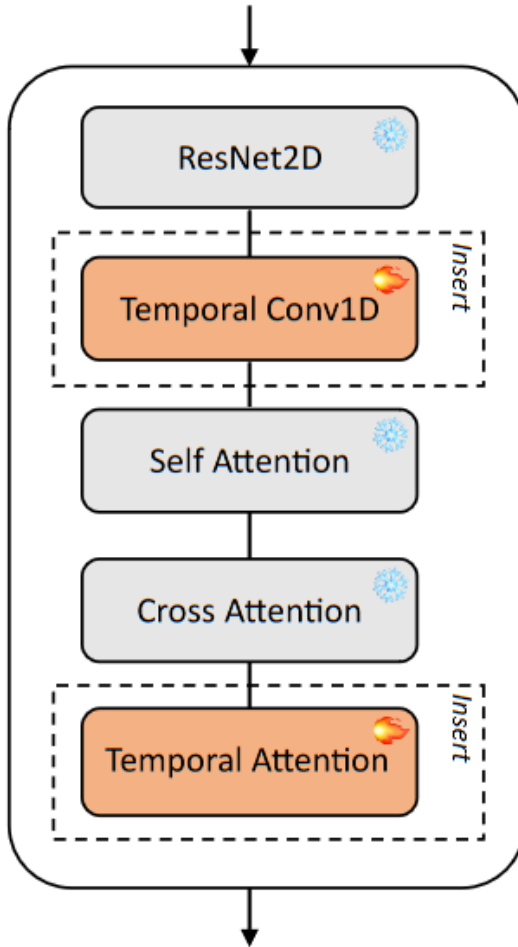
Part 1- Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs) and U-Net architecture for text to image model.



Method

Part 2 -Turn image U-Net to video



To endow the model with temporal understanding and produce coherent frames, Show-1 integrate temporal layers within each U-Net block.

After every Resnet2D block, we introduce a temporal convolution layer consisting of four 1D convolutions across the temporal dimension.

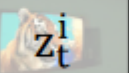
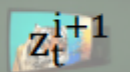
Following each self and cross-attention layer, we implement a temporal attention layer to facilitate dynamic temporal data assimilation.

Method

Part 3 – Pixel-based keyframe generation model

Given a text input, we initially produce a sequence of keyframes using a pixel-based Video U-Net at a very low spatial and temporal resolution.

Part 4 – Temporal interpolation diffusion module

Interpolation				
<i>noise</i>	x_t^i			x_t^{i+1}
<i>condition</i>		0	0	
<i>condition mask</i>	1	0	0	1

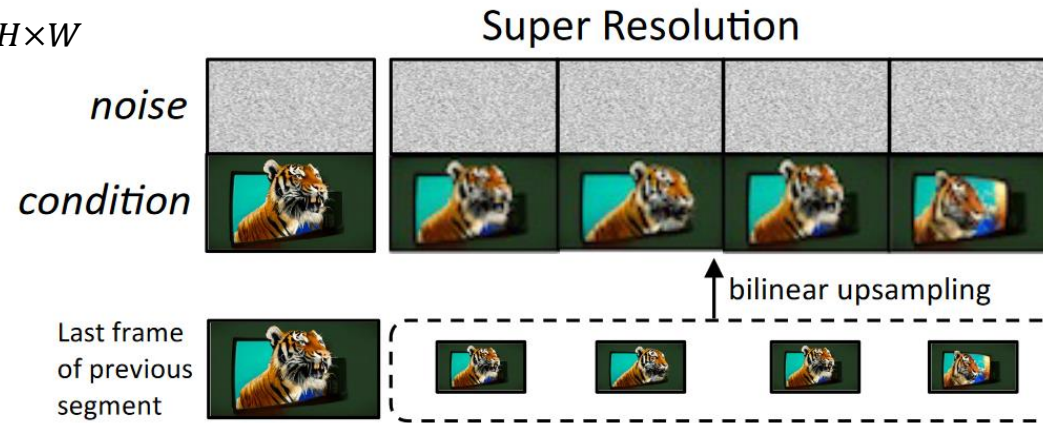
To enhance the temporal resolution of videos, they suggest a pixel-based temporal interpolation diffusion module. This method iteratively interpolates between the frames produced by Their keyframe modules.

Specially, they employ the **masking-conditioning mechanism**. Moreover, this module implements noise conditioning augmentation for z_t^i and z_t^{i+1} , which minimizes the vulnerability to domain disparities between the output from one cascade phase and the training inputs of the following phase.

Method

Part 5 – Super-resolution at low spatial resolution

Donate as $x' \in \mathcal{R}^{4T \times C \times H \times W}$
to $x'' \in \mathcal{R}^{4T \times C \times 4H \times 4W}$



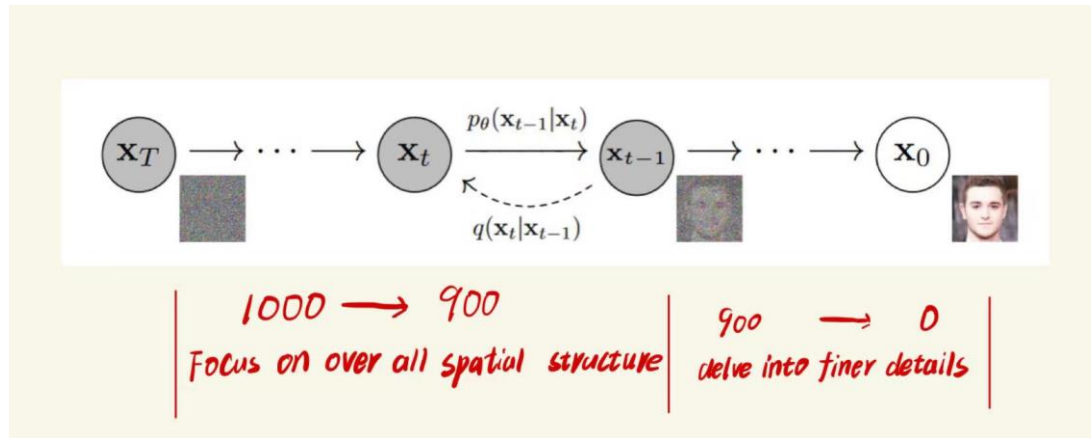
To improve the spatial quality of the videos, we introduce a pixel super-resolution approach utilizing the video U-Net. For this enhanced spatial resolution, we also incorporate **three additional channels**, which are populated using a bilinear upsampled low-resolution video clip.

To upscale all the interpolated frames on a standard GPU with 24G memory, divide the frames into four smaller segments and upscale each one individually. In the end, takes the upsampled last frame of one segment to complete the three supplementary channels of the initial frame in the following segment.

Method

Part 6 – Super-resolution at high spatial resolution

Various diffusion steps assume distinct roles during the generation process.



Therefore, a special U-Net only the 0 to 900 timesteps is trained, which can significantly enhances the end video quality, making this model a expert emphasizing high-resolution nuances, namely **expert translation**.

Experiment

Part 1- Implementation details

For the generation of pixel-based keyframes, we utilized **DeepFloyd1** as our pre-trained Text-to-Image model for initialization, producing videos of dimensions $8 \times 64 \times 40 \times 3 (T \times H \times W \times 3)$.



In our interpolation model, we initialize the weights using the keyframes generation model and produce videos with dimensions of $29 \times 64 \times 40 \times 3$.



For our initial model, we employ **DeepFloyd's SR** model for spatial weight initialization, yielding videos of size $29 \times 256 \times 160$.



In the subsequent super-resolution model, we modify the **ModelScope** text-to-video model and use our proposed expert translation to generate videos of $29 \times 576 \times 320$.

Experiment

Part 2- Quantitative results

UCF-101 Experiment

Table 1: Zero-shot text-to-video generation on UCF-101. Our approach achieves competitive results in both inception score and FVD metrics.

Method	IS (\uparrow)	FVD (\downarrow)
CogVideo (Hong et al., 2022) (English)	25.27	701.59
Make-A-Video (Singer et al., 2022)	<u>33.00</u>	367.23
MagicVideo (Zhou et al., 2022)	-	655.00
Video LDM (Blattmann et al., 2023a)	33.45	550.61
VideoFactory (Wang et al., 2023b)	-	410.00
Show-1 (ours)	35.42	<u>394.46</u>

UCF101 stands out as a categorized video dataset curated for action recognition tasks.

Experiment

Part 2- Quantitative results

MSR-VTT Experiment

Table 2: Quantitative comparison with state-of-the-art models on MSR-VTT. Our approach achieves the state-of-the-art performance.

Models	FID-vid (\downarrow)	FVD (\downarrow)	CLIPSIM (\uparrow)
NÜWA (Wu et al., 2022a)	47.68	-	0.2439
CogVideo (Chinese) (Hong et al., 2022)	24.78	-	0.2614
CogVideo (English) (Hong et al., 2022)	23.59	1294	0.2631
MagicVideo (Zhou et al., 2022)	-	1290	-
Video LDM (Blattmann et al., 2023a)	-	-	0.2929
Make-A-Video (Singer et al., 2022)	13.17	-	0.3049
ModelScopeT2V (Wang et al., 2023a)	11.09	550	0.2930
Show-1 (ours)	13.08	538	0.3072

The MSR-VTT dataset test subset comprises 2, 990 videos, accompanied by 59, 794 captions.

It's noteworthy that the pixel-based part of Show-1 are solely trained on the publicly available WebVid-10M dataset, in contrast to the Make-A-Video models, which are trained on other data.

Experiment

Part 2- Quantitative results

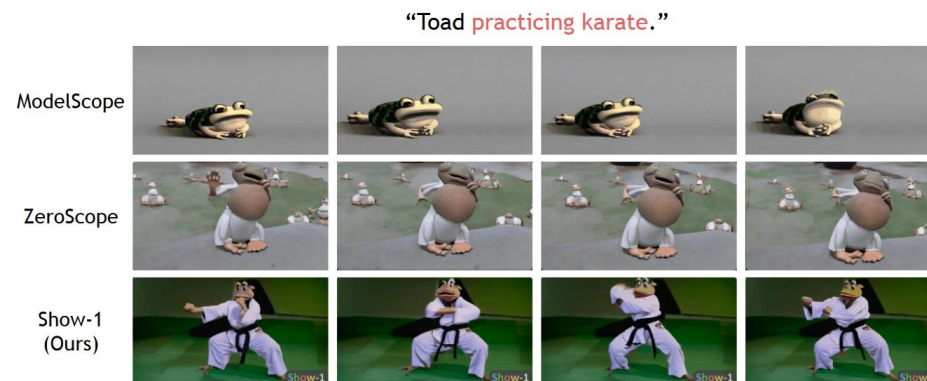
Human evaluation

Table 3: Human evaluation on state-of-the-art open-sourced text-to-video models.

	Video Quality	Text-Video alignment	Motion Fidelity
Ours <i>vs.</i> ModelScope	62% <i>vs.</i> 38%	63% <i>vs.</i> 37%	63% <i>vs.</i> 37%
Ours <i>vs.</i> ZeroSope	62% <i>vs.</i> 38%	58% <i>vs.</i> 42%	59% <i>vs.</i> 41%

Show-1 achieves the best human preferences on all evaluation parts.

Part 3- Quantitative results



Show-1 exhibits superior text-video alignment and visual fidelity compared to other recently open-sourced models.

Experiment

Part 4- Ablation studies

Table 4: Comparisons of different combinations of pixel-based and latent-based VDMs in terms of text-video similarity and memory usage during inference.

Low resolution stage	High resolution stage	CLIPSIM	Max memory
latent-based	latent-based	0.2934	15GB
latent-based	pixel-based	–	72GB
pixel-based	pixel-based	–	72GB
pixel-based	latent-based	0.3072	15GB

Impact of different combinations of pixel-based and latent-based VDMs



Figure 6: Qualitative comparison for our expert translation. With expert translation, the visual quality is significantly improved.

Impact of expert translation of latent-based VDM as super-resolution model

Demo Results



A burning lamborghini driving on rainbow.



A panda besides the waterfall is holding a sign that says "Show Lab".



Snail slowly creeping along super macro close-up, high resolution, best quality.



A blue tiger in the grass in the sunset, surrounded by butterflies.

Thanks!