# Factorizing Text-to-Video Generation by Explicit Image Conditioning

Rohit Girdhar[†,*], Mannat Singh[†,*], Andrew Brown[*], Quentin Duval[*], Samaneh Azadi[*], Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, Ishan Misra[*]

https://emu-video.metademolab.com/

GenAI, Meta

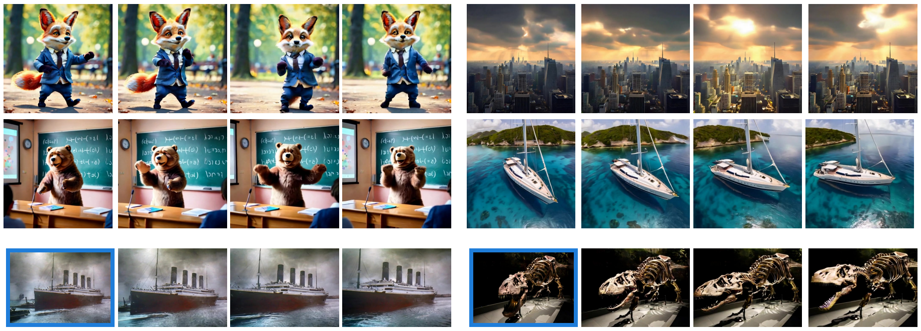


**Fig. 1:** **Emu Video** can generate high quality and temporally consistent videos while using a text prompt as input (top two rows), or an additional user-provided image (bottom row). Prompts: (top-left) A fox dressed in a suit dancing in a park, (top-right) The sun breaks through the clouds from the heights of a skyscraper, (middle-left): A bear is giving a presentation in the classroom, (middle-right): A 360 shot of a sleek yacht sailing gracefully through the crystal-clear waters of the Caribbean, (bottom-left): A ship driving off the harbor, (bottom-right): The dinosaur slowly comes to life. In the bottom two examples, a user-image is provided as an additional conditioning (shown in a blue border) and brought to life by Emu Video. The first one is a historical picture of the RMS Titanic departing from Belfast, Northern Ireland; and the second is a picture of a Tyrannosaurus rex fossil. Please see the website linked above for videos.

**Abstract.** We present Emu Video, a text-to-video generation model that factorizes the generation into two steps: first generating an image conditioned on the text, and then generating a video conditioned on the text and the generated image. We identify critical design decisions–adjusted noise schedules for diffusion, and multi-stage training–that enable us to directly generate high quality and high resolution videos, without requiring a deep cascade of models as in prior work. In human evaluations, our generated videos are strongly preferred in quality compared

[†] Equal first authors [*] Equal technical contribution

to all prior work–81% vs. Google's Imagen Video, 90% vs. Nvidia's PYOCO, and 96% vs. Meta's Make-A-Video. Our model outperforms commercial solutions such as RunwayML's Gen2 and Pika Labs. Finally, our factorizing approach naturally lends itself to animating images based on a user's text prompt, where our generations are preferred 96% over prior work.

## 1 Introduction

Large text-to-image models [17,21,28,35,55,62] trained on web-scale image-text pairs generate diverse and high quality images. While these models can be further adapted for text-to-video (T2V) generation [6,30,35,41,68] by using video-text pairs, video generation still lags behind image generation in terms of quality and diversity. Compared to image generation, video generation is more challenging as it requires modeling a higher dimensional spatiotemporal output space while still being conditioned only on a text prompt. Moreover, video-text datasets are typically an order of magnitude smaller than image-text datasets [17,35,68].

The dominant paradigm in video generation uses diffusion models [35, 68] to generate all video frames at once. In stark contrast, in NLP, long sequence generation is formulated as an autoregressive problem [11]: predicting one word conditioned on previously predicted words. Thus, the conditioning signal for each subsequent prediction progressively gets stronger. We hypothesize that strengthening the conditioning signal is also important for high quality video generation, which is inherently a time-series. However, autoregressive decoding with diffusion models [75] is challenging since generating a single frame from such models itself requires many iterations.

We propose Emu Video to strengthen the conditioning for diffusion based text-to-video generation with an explicit intermediate image generation step. Specifically, we factorize text-to-video generation into two subproblems: (1) generating an image from an input text prompt; (2) generating a video based on the stronger conditioning from the image *and* the text. Intuitively, giving the model a starting image and text makes video generation easier since the model only needs to predict how the image will evolve in the future.

Since video-text datasets are much smaller than image-text datasets, we initialize [6,68] our T2V model using a pretrained text-to-image (T2I) model whose weights are frozen. Unlike direct T2V methods [35,68], at inference, our factorized approach explicitly generates an image, allowing us to easily retain the visual diversity, style, and quality of the text-to-image model (see Figure 1). This allows Emu Video to outperform direct T2V methods, even when accounting for the same amount of training data, compute, and trainable parameters.

**Contributions.** We show that text-to-video (T2V) generation quality can be greatly improved by factorizing the generation into first generating an image and using the generated image and text to generate a video. We identify critical design decisions–changes to the diffusion noise schedule and multi-stage training–to efficiently generate videos at a high resolution of 512px bypassing the need for

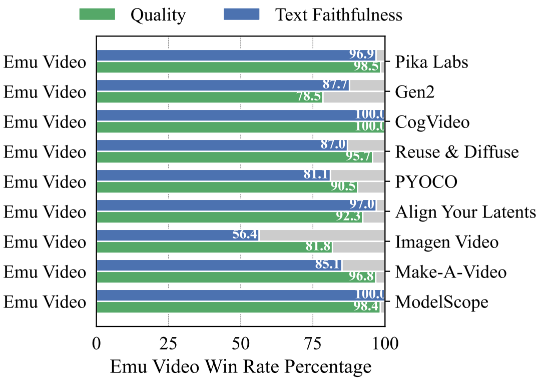


**Fig. 2: EMU VIDEO *vs.* prior work** in text-to-video in terms of video quality and text faithfulness win-rates evaluated by majority score of human evaluator preferences. Since most models from prior work are not accessible, we use the videos released by each method and their associated text prompt. The released videos are likely the *best* generations and we compare without any cherry-picking of our own generations. We also compare to commercial solutions (Gen2 [54] and PikaLabs [47]) and the open source model CogVideo [41] and ModelScope [77] using the prompt set from [6]. EMU VIDEO significantly outperforms all prior work across both metrics.

a deep cascade of models used in prior work [35,68]. We design a robust human evaluation scheme–JUICE–where we ask evaluators to justify their choices when making the selection in the pairwise comparisons of video generations. Figure 2 shows that EMU VIDEO significantly *surpasses all prior work* including commercial solutions: an average win rate of 91.8% for quality and 86.6% for text faithfulness. Beyond T2V, EMU VIDEO can be used out-of-the-box for image-to-video where the model generates a video based on a user-supplied image and a text prompt. In this setting, EMU VIDEO's generations are preferred 96% of the times over VideoComposer [78].

## 2 Related Work

**Text-to-Image (T2I) diffusion models.** Diffusion models [69] are a state-of-the-art approach for T2I generation, and out-perform prior GAN [8,43,66] or auto-regressive methods [1,23,29,60]. Diffusion models learn a data distribution by gradually denoising a normally distributed variable, often called 'noise', to generate the output. Prior work either denoises in the pixel space with pixel diffusion models [19,36,37,56,59,63], or in a lower-dimensional latent space with latent diffusion models [17,62]. In this work, we leverage latent diffusion models for video generation.

**Video generation/prediction.** Many prior works target the constrained settings of unconditional generation, or video prediction [45, 46, 53]. These approaches include training VAEs [4,5,18], auto-regressive models [25,41,42,61,83],

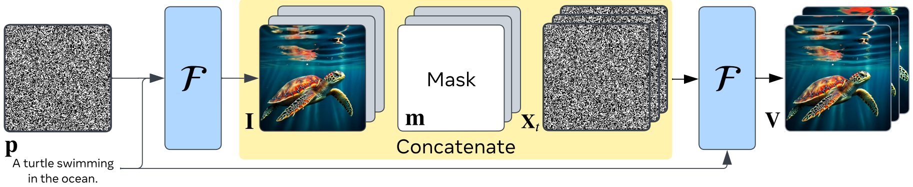


**Fig. 3: Factorized text-to-video generation** involves first generating an image $\mathbf{I}$ conditioned on the text $\mathbf{p}$, and then using stronger conditioning–the generated image *and* text–to generate a video $\mathbf{V}$. To condition our model $\mathcal{F}$ on the image, we zero-pad the image temporally and concatenate it with a binary mask indicating which frames are zero-padded, and the noised input.

masked prediction [27, 32, 88], LSTMs [67, 79], or GANs [2, 9, 16, 76]. However, these approaches are trained/evaluated on limited domains. In this work, we target the broad task of open-set T2V generation.

**Text-to-Video (T2V) generation.** Most prior works tackle T2V generation by leveraging T2I models. Several works take a training-free approach [40, 44, 49, 90] for *zero-shot* T2V generation by injecting motion information in the T2I models. Tune-A-Video [81] targets *one-shot* T2V generation by fine-tuning a T2I model with a single video. While these methods require no or limited training, the quality and diversity of the generated videos is limited.

Many prior works instead improve T2V generation by learning a *direct mapping* from the text condition to the generated videos by introducing temporal parameters to a T2I model [6, 30, 33, 39, 41, 48, 72, 74, 75, 80, 84, 86]. Make-A-Video [68] utilizes a pre-trained T2I model [59] and the prior network of [59] to train T2V generation without paired video-text data. Imagen Video [35] builds upon the Imagen T2I model [63] with a cascade of diffusion models [37, 39]. To address the challenges of modeling the high-dimensional spatiotemporal space, several works instead train T2V diffusion models in a lower-dimensional latent space [3,6,24,30,31,34,82], by adapting latent diffusion T2I models. Blattmann *et al*. [6] freeze the parameters of a pre-trained T2I model and train new temporal layers, whilst Ge *et al*. [30] build on [6] and design a noise prior tailored for T2V generation. The limitation of these approaches is that learning a direct mapping from text to the high dimensional video space is challenging. We instead strengthen our conditioning signal by taking a factorization approach. Unlike prior work that enhancing the conditions for T2V generation including leveraging large language models (LLMs) to improve textual description and understanding [24,40,50], or adding temporal information as conditions [14,78,85,90], our method does not require any models to generate the conditions as we use the first frame of a video as the image condition.

**Factorized generation.** The most similar works to Emu Video, in terms of factorization, is CogVideo [41] and Make-A-Video [68]. CogVideo builds upon the pretrained T2I model [20] for T2V generation using auto-regressive Transformer. The auto-regressive nature is fundamentally different to our explicit image conditioning in both training and inference stages. Make-A-Video [68] leverages the image embedding condition learnt from a shared image-text space. Our factorization leverage the first frame as is, which is a stronger condition. Moreover, Make-A-Video initializes from a pretrained T2I model but finetunes all the parameters so it cannot retain the visual quality and diversity of the T2I model as we do. Stable Video Diffusion [7] is a concurrent work that introduces similar factorization as ours for T2V generation.

# 3 Approach

The goal of text-to-video (T2V) generation is to construct a model that takes as input a text prompt $\mathbf{p}$ to generate a video $\mathbf{V}$ consisting of $T$ RGB frames. Recent methods [6,30,35,68] directly generate the $T$ video frames at once using text-only conditioning. Our approach builds on the hypothesis that stronger conditioning by way of both text *and* image can improve video generation (*cf*. § 3.2).

## 3.1 Preliminaries

Conditional Diffusion Models [36, 69] are a class of generative models that are trained to generate the output using a conditional input $\mathbf{c}$ by iteratively denoising from gaussian noise. At training time, time-step $t \in [0, N]$ dependent gaussian noise $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is added to the original input signal $\mathbf{X}$ to obtain a noisy input $\mathbf{X}_t = \alpha_t \mathbf{X} + \sqrt{1 - \alpha_t}\boldsymbol{\epsilon}_t$. $\alpha_t$ defines the "noise schedule", *i.e.*, noise added at timestep $t$ and $N$ is the total number of diffusion steps. The diffusion model is trained to denoise $\mathbf{X}_t$ by predicting either $\boldsymbol{\epsilon}_t$, $\mathbf{X}$, or $v_t = \alpha_t \boldsymbol{\epsilon}_t - \sqrt{1 - \alpha_t}\mathbf{X}$ (called v-prediction [65]). The signal-to-noise ratio (SNR) at timestep $t$ is given by $(\frac{\alpha_t}{1-\alpha_t})^2$ and decreases as $t \rightarrow N$. At inference, samples are generated by starting from pure noise $\mathbf{X}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and denoising it. Note that at inference time $\mathbf{X}_N$ has no signal, *i.e.*, zero SNR which has significant implications for video generation as we describe in § 3.2.

## 3.2 Emu Video

We factorize text-to-video generation into two steps (1) generating the first frame (image) given the text prompt $\mathbf{p}$ and (2) generating $T$ frames of a video by leveraging the text prompt and the image conditioning. We implement both steps using a latent diffusion model $\mathcal{F}$, illustrated in Figure 3. We initialize $\mathcal{F}$ with a pre-trained text-to-image model to ensure that it is capable of generating images at initialization. Thus, we only need to train $\mathcal{F}$ to solve the second step, *i.e.*, extrapolate a video conditioned on a text prompt and a starting frame. We train $\mathcal{F}$ using video-text pairs by sampling a starting frame $\mathbf{I}$ and asking the model to predict the $T$ frames using both the text prompt $\mathbf{p}$ and the image $\mathbf{I}$ conditioning. We denote a video $\mathbf{V}$ consisting of $T$ RGB frames of spatial dimensions $H', W'$ as a 4D tensor of shape $T \times 3 \times H' \times W'$. Since we use latent diffusion models, we first convert the video $\mathbf{V}$ into a latent space $\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W}$ using a image autoencoder applied frame-wise, which reduces the spatial dimensions. The latent space can be converted back to the pixel space using the autoencoder's decoder. The $T$ frames of the video are noised independently to produce the noised input $\mathbf{X}_t$, which the diffusion model is trained to denoise.

**Image conditioning.** We condition on the starting frame, $\mathbf{I}$, by concatenating it with the noise. Our design allows the model to use all the information in $\mathbf{I}$ unlike [68,78] that lose image information by conditioning on a semantic embedding. We represent $\mathbf{I}$ as a single-frame video ($T = 1$) and zero-pad it to obtain a $T \times C \times H \times W$ tensor. We use a binary mask $\mathbf{m}$ of shape $T \times 1 \times H \times W$ that is set to 1 at the first temporal position to indicate the position of the starting frame, and zero otherwise. The mask $\mathbf{m}$, starting frame $\mathbf{I}$, and the noised video $\mathbf{X}_t$ are concatenated channel-wise as the input to the model.

**Model.** We initialize our latent diffusion model $\mathcal{F}$ using the pretrained T2I model [17]. Like prior work [68], we add new learnable temporal parameters: a 1D temporal convolution after every spatial convolution, and a 1D temporal attention layer after every spatial attention layer. The original spatial convolution and attention layers are applied to each of the $T$ frames independently and are kept frozen. The pretrained T2I model is text conditioned and combined with the image conditioning (above), $\mathcal{F}$ is conditioned on both text and image.

**Zero terminal-SNR noise schedule.** We found that the diffusion noise schedules used in prior work [17,62] have a train-test discrepancy which prevents high quality video generation (reported for images in [13,51]). At training, the noise schedule leaves some residual signal, *i.e.*, has non-zero signal-to-noise (SNR) ratio even at the terminal diffusion timestep $N$. This prevents the diffusion model from generalizing at test time when we sample from random gaussian noise with no signal about real data. The residual signal is higher for high resolution video frames, due to redundant pixels across both space and time. We resolve this issue by scaling the noise schedule and setting the final $\alpha_N = 0$ [51], which leads to zero SNR at the terminal timestep $N$ during training too. We find that this design decision is *critical* for high resolution video generation.

**Interpolation model.** We use an interpolation model $\mathcal{I}$, architecturally the same as $\mathcal{F}$, to convert a low frame-rate video of $T$ frames into a high frame-rate video of $T_p$ frames. The interpolation model operates on $T_p \times C \times H \times W$ inputs/outputs. For frame conditioning, the input $T$ frames are zero-interleaved to produce $T_p$ frames, and a binary mask $\mathbf{m}$ indicating the presence of the $T$ frames are concatenated to the noised input (similar to the image conditioning for $\mathcal{F}$). The model is trained on video clips of $T_p$ frames of which $T$ frames are fed as input. For efficiency, we initialize $\mathcal{I}$ from $\mathcal{F}$ and only train the temporal parameters of the model $\mathcal{I}$ for the interpolation task.

**Simplicity in implementation.** Emu Video can be trained using standard video-text datasets, and does not require a deep cascade of models, *e.g.*, 7 models in [35], for generating high resolution videos. At inference, given a text prompt, we run $\mathcal{F}$ without the temporal layers to generate an image $\mathbf{I}$. We then use $\mathbf{I}$ and the text prompt as input to $\mathcal{F}$ to generate $T$ video frames, directly at high resolution. We can increase the fps of the video using $\mathcal{I}$. Since the spatial layers are initialized from a pretrained T2I model and kept frozen, our model retains the conceptual and stylistic diversity learned from large image-text datasets, and uses it to generate $\mathbf{I}$. This comes at no additional training cost unlike [35] that jointly finetune on image and video data to maintain such style. Many direct T2V approaches [6,68] also initialize from a pretrained T2I model and keep the spatial layers frozen. However, they do not employ our image-based factorization failing to retain the quality and diversity in the T2I model.

**Robust human evaluation (JUICE).** Similar to recent studies [17,35,57,68], we find that the automatic evaluation metrics [73] do not reflect improvements in quality. We primarily use human evaluation to measure T2V generation performance on two orthogonal aspects - (a) video generation quality denoted as Quality (Q) and (b) the alignment or 'faithfulness' of the generated video to the text prompt, denoted as Faithfulness (F). We found that asking human evaluators to JUstify their choICE (JUICE) when picking a generation over the other significantly improves the inter-annotator agreement (details in Sec. 3). The annotators select one or more pre-defined reasons to justify their choice. The reasons for picking one generation over the other for Quality are: pixel sharpness, motion smoothness, recognizable objects/scenes, frame consistency, and amount of motion. For Faithfulness we use two reasons: spatial text alignment, and temporal text alignment.

### 3.3 Implementation Details

We provide complete implementation details in the supplement Sec. 1 and highlight salient details next.

**Architecture and initialization.** We adapt the text-to-image U-Net architecture from [17] for our model and initialize all the spatial parameters with the pretrained model. The pretrained model produces square 512px images using an 8 channel $64 \times 64$ latent as the autoencoder downsamples spatially by $8\times$. The model uses both a frozen T5-XL [15] and a frozen CLIP [58] text encoder to extract features from the text prompt. Separate cross-attention layers in the U-Net attend to each of the text features. After initialization, our model contains 2.7B frozen spatial parameters, and 1.7B trainable temporal parameters.

The temporal parameters are initialized as identity operations: identity kernels for convolution, and zeroing the final MLP layer of the temporal attention block. In our preliminary experiments, the identity initialization improved the model convergence by $2\times$. For the additional channels in the model input due to image conditioning, we add $C+1$ additional learnable channels (zero-initialized) to the kernel of the first spatial convolution layer. Our model produces 512px square videos of $T = 8$ or 16 frames and is trained with square center-cropped video clips of 1, 2 or 4 seconds sampled at 8fps or 4fps. We train all our models with a batch size of 512 and describe the details next.

**Efficient multi-stage multi-resolution training.** To reduce the computational complexity, we train in two stages - (1) for majority of the training iterations (70K) we train for a simpler task: 256px 8fps 1s videos, which reduces

| Method | Q | F |
|---|---|---|
| Factorized | 70.5 | 63.3 |

(a)

| Method | Q | F |
|---|---|---|
| Zero SNR | 96.8 | 88.3 |

(b)

| Method | Q | F |
|---|---|---|
| Multi-stage | 81.8 | 84.1 |

(c)

| Method | Q | F |
|---|---|---|
| HQ finetuned | 65.1 | 79.6 |

(d)

| Method | Q | F |
|---|---|---|
| Frozen spatial | 55.0 | 58.1 |

(e)

**Table 1: Key design decisions in Emu Video**. Each table shows the preference, in terms of the Quality (Q) and Faithfulness (F), on adopting a design decision *vs*. a model that does not have it. Our results show clear preference to a) factorized generation that uses both image and text conditioning (against a direct video generation baseline that is only text conditioned), b) adopting zero terminal-SNR noise schedule for directly generating high resolution 512px videos, c) adopting the multi-stage training setup compared to training directly at the high resolution, d) incorporating the high quality (HQ) finetuning, and e) freezing the spatial parameters. See § 4.1 for details.

per-iteration time by 3.5× due to the reduction in spatial resolution; (2) we then train the model at the desired 512px resolution on 4fps 2s videos for 15K iterations. The change in spatial resolution does not affect the 1D temporal layers. Although the frozen spatial layers were pretrained at 512px, changing the spatial resolution at inference to 256px led to no loss in generation quality. We use the noise schedule from [62] for 256px training, and with zero terminal-SNR for 512px training using the v-prediction objective [65] with $N = 1000$ steps for the diffusion training. We sample from our models using 250 steps of DDIM [70]. Optionally, to increase duration, we further train the model on 16 frames from a 4s video clip for 25K iterations.

**Finetuning for higher quality.** Similar to the observation in image generation [17], we find that the motion of the generated videos can be improved by finetuning the model on a small subset of high motion and high quality videos. We automatically identify a small finetuning subset of 1.6K videos from our training set which have high motion (computed using motion signals stored in `H.264` encoded videos). We follow standard practice [62] and also apply filtering based on aesthetic scores [62] and CLIP [58] similarity between the video's text and first frame. Specifically, we use a video with $N$ frames $\{f_j\}$ if $\text{CLIP}(f_1) > 0.25$, $\text{aesthetic}(f1) > 5.7$, $\min_{j=1}^{N-5} \sum_{i=j}^{j+5}(\text{motion score}(f_i)) > 0.5$.

**Interpolation model.** We initialize the interpolation model from the video model $\mathcal{F}$. Our interpolation model takes 8 frames as input and outputs $T_p = 37$ frames at 16fps. During training, we use noise augmentation [37] where we add noise to the frame conditioning by randomly sampling timesteps $t \in \{0, ...250\}$. At inference time, we noise augment the samples from $\mathcal{F}$ with $t = 100$.

## 4 Experiments

**Dataset.** We train Emu Video on a dataset of 34M licensed video-text pairs Our videos are 5-60 seconds long and cover a variety of natural world concepts. The videos were not curated for a particular task and were *not* filtered for text-frame similarity or aesthetics. Unless noted, we train the model on the full set, and do not use the 1.6K high motion quality finetuning subset described in § 3.3.

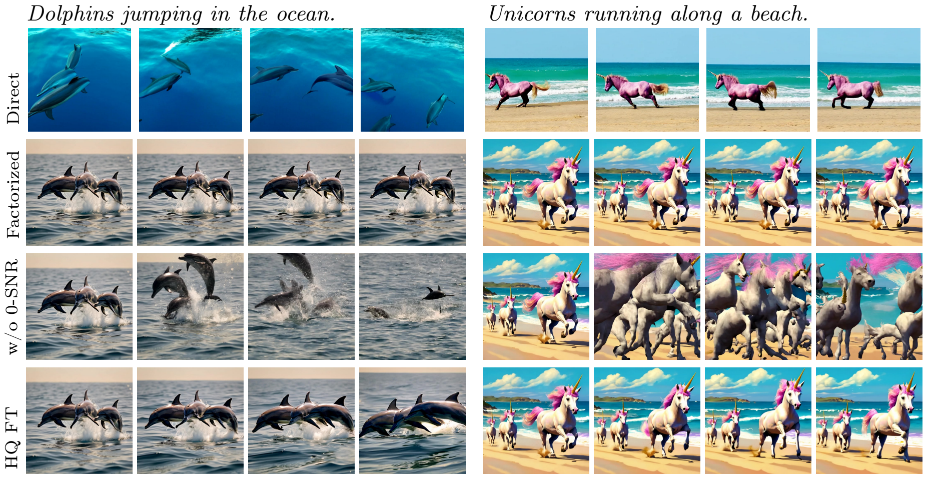

*Dolphins jumping in the ocean.*
*Unicorns running along a beach.*
Direct
Factorized
w/o 0-SNR
HQ FT

**Fig. 4: Design choices in Emu Video.** *Top row:* Direct text-to-video generation produces videos that have low visual quality and are inconsistent. *Second row:* We use a factorized text-to-video approach that produces high quality videos and improves consistency. *Third row:* Not using a zero terminal-SNR noise schedule at 512px generation leads to significant inconsistencies in the generations. *Bottom row:* Finetuning our model (second row) with HQ data increases the motion in the generated videos.

**Text prompt sets for human evaluation.** We use the text prompt sets from prior work (*cf*. Appendix Table 7) to generate videos. The prompts cover a wide variety of categories that can test our model's ability to generate natural and fantastical videos, and compose different visual concepts. We use our proposed JUICE evaluation scheme ( Sec. 3) for reliable human evaluation and use the majority vote from 5 evaluators for each comparison.

### 4.1 Ablating design decisions

We study the effects of our design decisions using the 8 frame generation setting and report human evaluation results in Table 1 using pairwise comparisons on the 307 prompt set of [68].

**Factorized *vs*. Direct generation.** We compare our factorized generation to a direct T2V generation model that generates videos from text condition only. We ensure that the pretrained T2I model, training data, number of training iterations, and trainable parameters are held constant for this comparison. As shown in Table 1a, the factorized generation model's results are strongly preferred both in Quality and Faithfulness.The strong preference in Quality is because the direct generation model does not retain the style and quality of the text-to-image model despite frozen spatial parameters, while also being less temporally consistent (examples in Figure 4).

**Zero terminal-SNR noise schedule.** We compare using zero terminal-SNR for the high resolution 512px training against a model that is trained with the standard noise schedule. Table 1b shows that generations using zero terminal-SNR are *strongly* preferred. This suggests that the zero terminal-SNR noise schedule's effect of correcting the train-test discrepancy as described in § 3.2 is critical for high resolution video generation. We also found that zero terminal-SNR has a stronger benefit for our factorized generation compared to a direct T2V model possibly. Similar to images [51], in the direct T2V case, this decision primarily affects the color composition. For our factorized approach, this design choice was critical for object consistency and high quality as our qualitative results in Figure 4 show.

**Multi-stage multi-resolution training.** We spend most training budget (4×) on the 256px 8fps stage compared to the 3.5× slower (due to increased resolution) 512px 4fps stage. We compare to a baseline that trains only the 512px stage with the same training budget. Table 1c shows that our multi-stage training yields significantly better results.

**High quality finetuning.** We study the effect of finetuning our model on automatically identified high quality videos in Table 1d. We found that this finetuning improves on both metrics, particularly the model's ability to respect the motion specified in the text prompt as reflected by the strong gain in Faithfulness.

**Parameter freezing.** We test if freezing the spatial parameters of our model affects performance by comparing it to a model where all parameters are finetuned during the second 512px training stage. For a fair comparison, the same conditioning images $\mathbf{I}$ are used across both models. Table 1e suggests that freezing the spatial parameters produces better videos, while reducing training cost.

### 4.2 Comparison to prior work

We evaluate Emu Video against prior work and train $\mathcal{F}$ to produce 16 frame 4 second long videos and use the best design decisions from § 4.1, including high quality finetuning. We use the interpolation model $\mathcal{I}$ on our generations to get 16fps videos. Please see Sec. 1 for details on how we interpolate 16-frame videos.

**Human evaluation of text-to-video generation.** Since many recent prior methods in text-to-video generation are closed source [6, 30, 31, 35], we use the publicly released examples from each of these methods. Note that the released videos per method are likely to be the 'best' representative samples from each method and may not capture their failure modes. For Make-A-Video, we obtained non cherry-picked generations through personal communication with the authors. For CogVideo [41], we perform T2V generation on the prompt set from [6] using the open source models. We also benchmark against commercially engineered black-box text-to-video solutions, Gen2 [54] and PikaLabs [47], obtaining generations through their respective websites using the prompts from [6]. We do not cherry-pick or contrastively rerank [60, 87] our videos, and generate them using a deterministic random noise seed that is not optimized in any way.

Since each method generates videos at different resolutions, aspect-ratios, and frame-rates, we reduce annotator bias in human evaluations by postprocessing the videos for each comparison in Figure 2 so that they match in these aspects. Full details on this postprocessing and the text prompts used are in Sec. 4. As

*Flying through an intense battle between pirate ships in a stormy ocean.*

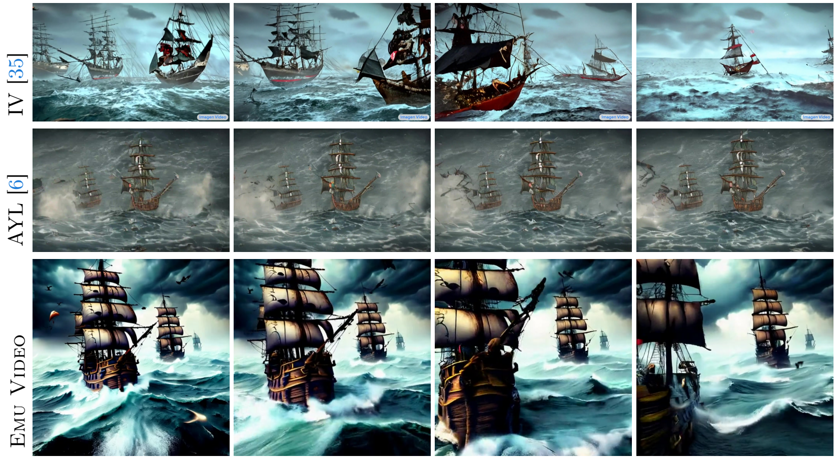


**Fig. 5: Qualitative comparison.** Emu Video produces higher quality generations compared to Imagen Video [35] and Align Your Latents [6] in terms of style and consistency.

shown in Figure 2, Emu Video's generations significantly outperform all prior work, including commercial solutions, both in terms of Quality (by an average of 91.8%) and Faithfulness (by an average of 86.6%). We show some qualitative comparisons in Figure 5 and some additional generations in Figure 1. Emu Video generates videos with significantly higher quality, and overall faithfulness to both the objects and motion specified in the text. Since our factorized approach explicitly generates an image, we retain the visual diversity and styles of the T2I model, leading to far better videos on fantastical and stylized prompts. Additionally, Emu Video generates videos with far greater temporal consistency than prior work. We hypothesize that since we use stronger conditioning of image and text, our model is trained with a relatively easier task of predicting how an image evolves into the future, and thus is better able to model the temporal nature of videos. Please see Sec. 5 for more qualitative comparisons. We include human evaluations where videos are not postprocessed in the supplement Sec. 4, where again Emu Video's generations significantly outperform all prior work. The closest model in performance compared to ours is Imagen Video when measured on Faithfulness, where we outperform Imagen Video by 56%. Imagen Video's released prompts ask for generating text characters, a known failure mode [17, 62] of latent diffusion models used in Emu Video.

We inspect the reasons that human evaluators prefer Emu Video generations over the two strongest competitors in Figure 6. A more detailed inspection is provided in Sec. 3. Emu Video generations are preferred due to their better pixel sharpness and motion smoothness. While being state-of-the-art, Emu Video is also simpler and has a two model cascade with a total of 6.0B parameters (2.7B frozen parameters for spatial layers, and 1.7B learnable temporal parameters

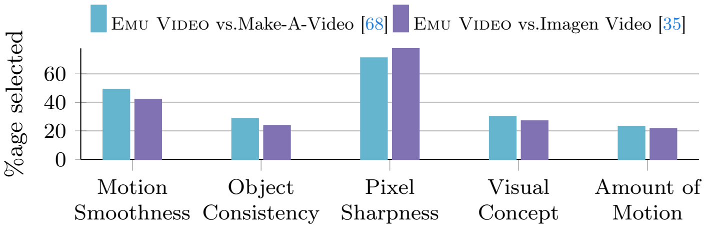


**Fig. 6: Percentage of each reason selected for samples where Emu Video wins against Make-A-Video [68] or Imagen Video [35] on Quality**. Human raters pick Emu Video primarily due to their pixel sharpness and motion smoothness, with an overall preference of 96.8% and 81.8% to each baseline, respectively.

each for $\mathcal{F}$ and $\mathcal{I}$), which is much simpler than methods like Imagen Video (7 model cascade, 11.6B parameters), Make-A-Video (5 model cascade, 9.6B parameters) trained using similar scale of data.

| **Method** | **Automated** | |
|---|---|---|
| | **FVD ↓** | **IS ↑** |
| MagicVideo [91] | 655.0 | - |
| Align Your Latents [6] | 550.6 | 33.5 |
| Make-A-Video [68] | 367.2 | 33.0 |
| PYOCO [30] | 355.2 | 47.8 |
| Emu Video | 317.1 | 42.7 |

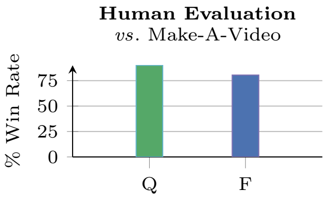


**Table 2: Automated metrics for zero-shot text-to-video evaluation on UCF101.** (Left) We present automated metrics and observe that Emu Video achieves competitive IS and outperforms all prior work on FVD. (Right) We conduct human evaluations to compare Emu Video and Make-A-Video where Emu Video significantly outperforms Make-A-Video both in Quality (90.1%) and Faithfulness (80.5%).

**Automated metrics.** In Table 2, we compare against prior work using the zero-shot T2V generation setting from [68] on the UCF101 dataset [71]. Emu Video achieves a comptetitive IS score [64] and a lower FVD [73]. To confirm these automated scores, we also use human evaluations to compare our generations to Make-A-Video. We use a subset of 303 generated videos (3 random samples per UCF101 class) and find that our generations are strongly preferred (Table 2 right). Qualitative comparisons can be found in Sec. 5.

**Animating images.** A benefit of our factorized generation is that the same model can be used out-of-the-box to 'animate' user-provided images by supplying them as the conditioning image **I**. We compare Emu Video's image animation with six methods, prior and concurrent work [7, 12, 78, 89] and commercial image-to-video (I2V) solutions [47, 54], on the prompts from [68] and [6]. All the methods are shown the same image generated using a different text-to-image model [57] and expected to generate a video according to the text prompt[⋆]. We

[⋆] Due to lack of access to training data of SDXL [57] and their underlying model, we leveraged their corresponding APIs for our comparison.

| **Method** | #Prompts | **Q** | **F** |
|---|---|---|---|
| Emu Video *vs.* VideoComposer I2V * [78] | 65 [6] | 96.9 | 96.9 |
| Emu Video *vs.* PikaLabs I2V * [47] | | 84.6 | 84.6 |
| Emu Video *vs.* Gen2 I2V * [54] | | 70.8 | 76.9 |
| Emu Video *vs.* VideoCrafter I2V * [12] | | 81.5 | 80.0 |
| Emu Video *vs.* Stable Video Diffusion I2V ** [7] | | 72.3 | 73.9 |
| Emu Video *vs.* I2VGen-XL I2V ** [89] | | 69.2 | 66.1 |
| Emu Video *vs.* VideoComposer I2V * [78] | 307 [68] | 97.4 | 91.2 |

**Table 3: Human evaluation of Emu Video *vs.* prior* and concurrent** work in text-conditioned image animation.** We compare Emu Video against six methods across two prompt sets using generations from [57] as the starting images. Emu Video's animated videos are strongly preferred over all baselines.

report human evaluations in Table 3 and automated metrics in the supplement (*cf*. Appendix Table 6). Human evaluators strongly prefer Emu Video's generations across all the baselines. These results demonstrate the superior image animation capabilities of Emu Video compared to methods specifically designed for the image-to-video task.

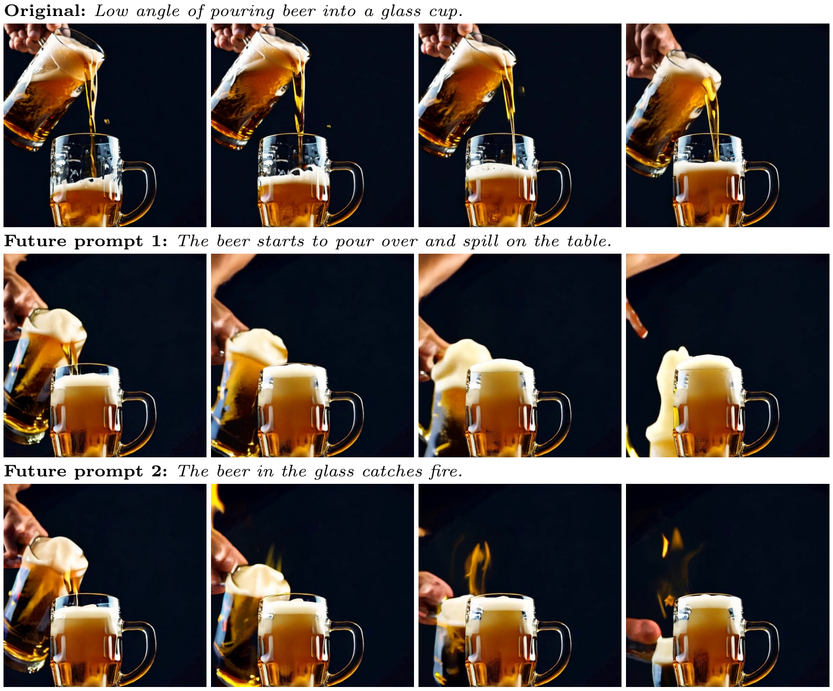
### 4.3 Analysis

**Nearest neighbor baseline.** We expect good and useful generative models to outperform a nearest neighbor retrieval baseline and create videos not in the training set. We construct a strong nearest neighbor baseline that retrieves videos from the full training set (34M videos) by using the text prompt's CLIP feature similarity to the training prompts. When using the evaluation prompts from [68], human evaluators prefer Emu Video's generations 81.1% in Faithfulness over real videos confirming that Emu Video outperforms the strong retrieval baseline. We manually inspected and confirmed that Emu Video outperforms the baseline for prompts not in the training set.

**Extending video length with longer text.** Recall that our model conditions on the text prompt and a starting frame to generate a video. With a small architectural modification, we can also condition the model on $T$ frames and *extend* the video. Thus, we train a variant of Emu Video to generate the future 16 frames conditioned on the 'past' 16 frames. While extending the video, we use a *future* text prompt different from the one used for the original video and visualize results in Figure 7. We find that the extended videos respect the original video as well as the future text prompt.

## 5 Limitations and ethical considerations

We presented Emu Video, a factorized approach to text-to-video generation that leverages strong image and text conditioning. Emu Video significantly outperforms all prior work including commercial solutions. Although our model has been a step change in video generation and shares valuable insights into the modeling and evaluation challenges, there are limitations. Emu Video can be improved in the following aspects as future research directions: the realism of the



**Original:** *Low angle of pouring beer into a glass cup.*

**Future prompt 1:** *The beer starts to pour over and spill on the table.*

**Future prompt 2:** *The beer in the glass catches fire.*

**Fig. 7: Extending to longer videos.** We test a variant of Emu Video that is conditioned on all the frames from the original video, and generates new videos conditioned on a future prompt. For two different future prompts, our model generates plausible extended videos that respect the original video and the future text.

presented content, fine-grained details such as hand and face artifacts, modeling physics, and maintaining quality and consistency for long video durations. These factors have been considered in the JUICE metric where the raters are asked to consider object/scene consistency and pixel quality in their evaluations. Another direction for future research is to improve Emu Video's ability to recover from conditioning frames that are not representative of the prompt. Strengthening the conditioning for video models using pure autoregressive decoding with diffusion models is not currently computationally attractive. However, further research may provide benefits for longer video generation.

***Ethical considerations.*** We propose advancements in generative methods specifically to improve the generation of high dimensional video outputs. Generative methods can be applied to a large variety of different usecases which are beyond the scope of this work. A careful study of the data, model, its intended applications, safety, risk, bias, and societal impact is necessary before any real world application.

**Acknowledgments.** We are grateful for the support of multiple collaborators at Meta who helped us in this work. Baixue Zheng, Baishan Guo, Jeremy Teboul, Milan Zhou, Shenghao Lin, Kunal Pradhan, Jort Gemmeke, Jacob Xu, Dingkang Wang, Samyak Datta, Guan Pang, Symon Perriman, Vivek Pai, Shubho Sengupta for their help with the data and infra. We would like to thank Uriel Singer, Adam Polyak, Shelly Sheynin, Yaniv Taigman, Licheng Yu, Luxin Zhang, Yinan Zhao, David Yan, Emily Luo, Xiaoliang Dai, Zijian He, Peizhao Zhang, Peter Vajda, Roshan Sumbaly, Armen Aghajanyan, Michael Rabbat, and Michal Drozdzal for helpful discussions. We are also grateful to the help from Lauren Cohen, Mo Metanat, Lydia Baillergeau, Amanda Felix, Ana Paula Kirschner Mofarrej, Kelly Freed, Somya Jain. We thank Ahmad Al-Dahle and Manohar Paluri for their support.

## References

1. Aghajanyan, A., Huang, P.Y.B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., Zettlemoyer, L.: Cm3: A causal masked multimodal model of the internet. ArXiv **abs/2201.07520** (2022)
2. Aldausari, N., Sowmya, A., Marcus, N., Mohammadi, G.: Video generative adversarial networks: A review. ACM Comput. Surv. **55**(2) (jan 2022). https://doi.org/10.1145/3487891, https://doi.org/10.1145/3487891
3. An, J., Zhang, S., Yang, H., Gupta, S., Huang, J.B., Luo, J., Yin, X.: Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation (2023)
4. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R.H., Levine, S.: Stochastic variational video prediction. In: ICLR (2018), https://openreview.net/forum?id=rk49Mg-CW
5. Babaeizadeh, M., Saffar, M.T., Nair, S., Levine, S., Finn, C., Erhan, D.: Fitvid: Overfitting in pixel-level video prediction. arXiv preprint arXiv:2106.13195 (2020)
6. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 22563–22575 (2023), https://api.semanticscholar.org/CorpusID:258187553
7. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
8. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=B1xsqj09Fm
9. Brooks, T., Hellsten, J., Aittala, M., Wang, T.C., Aila, T., Lehtinen, J., Liu, M.Y., Efros, A.A., Karras, T.: Generating long videos of dynamic scenes. In: NeurIPS (2022)
10. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023)
11. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. preprint arXiv:2005.14165 (2020)

12. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv:2310.19512 (2023)
13. Chen, T.: On the importance of noise scheduling for diffusion models. arXiv preprint arXiv:2301.10972 (2023)
14. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
15. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
16. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets (2019)
17. Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)
18. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1174–1183. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/denton18a.html
19. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis (2021)
20. Ding, M., Zheng, W., Hong, W., Tang, J.: Cogview2: Faster and better text-to-image generation via hierarchical transformers. NeurIPS (2022)
21. Donahue, J., Krahenbühl, P., Darrell, T.: Adversarial feature learning. In: ICLR (2016)
22. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models (2023)
23. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
24. Fei, H., Wu, S., Ji, W., Zhang, H., Chua, T.S.: Empowering dynamics-aware text-to-video diffusion with large language models (2023)
25. Finn, C., Goodfellow, I., Levine, S.: Unsupervised learning for physical interaction through video prediction. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 64–72. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
26. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement **33**(3), 613–619 (1973)
27. Fu, T.J., Yu, L., Zhang, N., Fu, C.Y., Su, J.C., Wang, W.Y., Bell, S.: Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In: CVPR. pp. 10681–10692 (June 2023)
28. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. arXiv preprint arXiv:2203.13131 (2022)
29. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-scene: Scene-based text-to-image generation with human priors. In: European Conference on Computer Vision (2022)
30. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models (2023)

31. Gu, J., Wang, S., Zhao, H., Lu, T., Zhang, X., Wu, Z., Xu, S., Zhang, W., Jiang, Y.G., Xu, H.: Reuse and diffuse: Iterative denoising for text-to-video generation (2023)
32. Gupta, A., Tian, S., Zhang, Y., Wu, J., Martín-Martín, R., Fei-Fei, L.: Maskvit: Masked visual pre-training for video prediction. In: ICLR (2023), https://openreview.net/forum?id=QAV2CcLEDh
33. Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., Wood, F.: Flexible diffusion modeling of long videos. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS. vol. 35, pp. 27953–27965. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/b2fe1ee8d936ac08dd26f2ff58986c8f-Paper-Conference.pdf
34. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2023)
35. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models (2022)
36. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. arXiv preprint arxiv:2006.11239 (2020)
37. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. arXiv preprint arXiv:2106.15282 (2021)
38. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
39. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS. vol. 35, pp. 8633–8646. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/39235c56aef13fb05a6adc95eb9d8d66-Paper-Conference.pdf
40. Hong, S., Seo, J., Hong, S., Shin, H., Kim, S.: Large language models are frame-level directors for zero-shot text-to-video generation (2023)
41. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers (2022)
42. Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K.: Video pixel networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 1771–1779. PMLR (06–11 Aug 2017), https://proceedings.mlr.press/v70/kalchbrenner17a.html
43. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
44. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439 (2023)
45. Kim, T., Ahn, S., Bengio, Y.: Variational Temporal Abstraction. Curran Associates Inc., Red Hook, NY, USA (2019)
46. Kumar, M., Babaeizadeh, M., Erhan, D., Finn, C., Levine, S., Dinh, L., Kingma, D.: Videoflow: A conditional flow-based model for stochastic video generation. In: ICLR (2020), https://openreview.net/forum?id=rJgUfTEYvH
47. Labs, P.: Pika labs. https://www.pika.art/
48. Laptev, I., Lindeberg, T.: Space-time interest points. In: ICCV (2003)

49. Lee, S., Kong, C., Jeon, D., Kwak, N.: Aadiff: Audio-aligned video synthesis with text-to-image diffusion (2023)
50. Lian, L., Shi, B., Yala, A., Darrell, T., Li, B.: Llm-grounded video diffusion models. arXiv preprint arXiv:2309.17444 (2023)
51. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. arXiv preprint arXiv:2305.08891 (2023)
52. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
53. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error (2016)
54. ML, R.: Gen2. https://research.runwayml.com/gen2
55. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
56. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models (2022)
57. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
58. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision (2021)
59. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
60. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation (2021)
61. Ranzato, M., Szlam, A., Bruna, J., Mathieu, M., Collobert, R., Chopra, S.: Video (language) modeling: a baseline for generative models of natural videos. ArXiv **abs/1412.6604** (2014), https://api.semanticscholar.org/CorpusID:17572062
62. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021)
63. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding (2022)
64. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. NeurIPS **29** (2016)
65. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022)
66. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. vol. abs/2301.09515 (2023)
67. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.k., WOO, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) NeurIPS. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf

68. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., Parikh, D., Gupta, S., Taigman, Y.: Make-a-video: Text-to-video generation without text-video data. In: ICLR (2023), https://openreview.net/forum?id=nJfylDvgzlq
69. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2256–2265. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/sohl-dickstein15.html
70. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020), https://arxiv.org/abs/2010.02502
71. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01 (2012)
72. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion (2023)
73. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation (2019)
74. Villegas, R., Babaeizadeh, M., Kindermans, P.J., Moraldo, H., Zhang, H., Saffar, M.T., Castro, S., Kunze, J., Erhan, D.: Phenaki: Variable length video generation from open domain textual descriptions. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=vOEXS39nOF
75. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: MCVD - masked conditional video diffusion for prediction, generation, and interpolation. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) NeurIPS (2022)
76. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain. pp. 613–621 (2016), https://proceedings.neurips.cc/paper/2016/hash/04025959b191f8f9de3f924f0940515f-Abstract.html
77. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
78. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018 (2023)
79. Wichers, N., Villegas, R., Erhan, D., Lee, H.: Hierarchical long-term video prediction without supervision. In: International Conference on Machine Learning (2018), https://api.semanticscholar.org/CorpusID:49193136
80. Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., Duan, N.: Godiva: Generating open-domain videos from natural descriptions. ArXiv **abs/2104.14806** (2021), https://api.semanticscholar.org/CorpusID:233476314
81. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: ICCV (2023)
82. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: Simda: Simple diffusion adapter for efficient video generation (2023)
83. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers (2021)
84. Yang, R., Srivastava, P., Mandt, S.: Diffusion probabilistic modeling for video generation. arXiv preprint arXiv:2203.09481 (2022)

85. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
86. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., Fu, J., Ming, G., Wang, L., Liu, Z., Li, H., Duan, N.: Nuwa-xl: Diffusion over diffusion for extremely long video generation (2023)
87. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 (2022)
88. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A., Yang, M.H., Hao, Y., Essa, I., Jiang, L.: Magvit: Masked generative video transformer. In: CVPR (2023), https://arxiv.org/abs/2212.05199
89. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. arXiv preprint arXiv:2311.04145 (2023)
90. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation (2023)
91. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models (2023)