Supplemental for Factorizing Text-to-Video Generation by Explicit Image Conditioning

1 Implementation Details

In this section we include details on the architectures and hyper-parameters used for training the models in the main paper, and on the use of multiple conditionings for classifier-free guidance. For both our text-to-video ($\mathcal{F}$) and interpolation ($\mathcal{I}$) models we train with the same U-Net architecture. We share the exact model configuration for our U-Net in Table 1, and the configuration for our 8-channel autoencoder in Table 2.

Setting	**Value**
`input_shape`	[17, T, 64, 64]
`output_shape`	[8, T, 64, 64]
`model_channels`	384
`attention_resolutions`	[4, 2, 1]
`num_res_blocks`	[3, 4, 4, 4]
`channel_multipliers`	[1, 2, 4, 4]
`use_spatial_attention`	True
`use_temporal_attention`	True
`transformer_config:`	
`d_head`	64
`num_layers`	2
`context_dim_layer_1`	768
`context_dim_layer_2`	2048

Table 1: U-Net architecture details. Our U-Net contains 4.3B total parameters, out of which 2.7B are initialized from our pretrained text-to-image model and kept frozen, resulting in 1.7B trainable parameters. T is the total frames produced by the model.

Table 3 shares the training hyperparameters we used for various stages of our training – 256px training, 512px training, High Quality finetuning, and frame interpolation. For inference, we use the DDIM sampler [70] with 250 diffusion steps. We use Classifier Free Guidance (CFG) [38] with w_{img} of 7.5 for image generation, and w_{img} of 2.0 and w_{txt} of 7.5 for both video generation and frame interpolation. We share more details about handling multiple conditionings for Classifier Free Guidance next.

Setting	Value
type	AutoencoderKL [62]
z_channels	8
in_channels	3
out_channels	3
base_channels	128
channel_multipliers	[1, 2, 4, 4]
num_res_blocks	2

Table 2: VAE architecture details. We use an image based VAE and apply it to videos frame-by-frame. Our VAE encoder downsamples videos spatially by 8×8 and produces 8 channel latents.

Multiple Conditionings for CFG. For video generation, our model receives two conditioning signals (image $\mathbf{I}$, text prompt $\mathbf{p}$), which we use in conjunction for Classifier Free Guidance [38]. Eq 1 lists the combined CFG equation we use.

$$\tilde{\mathbf{X}} = \mathbf{X} + w_i(\mathbf{X}(\mathbf{I}) - \mathbf{X}(\varnothing)) + w_p(\mathbf{X}(\mathbf{I}, \mathbf{p}) - \mathbf{X}(\mathbf{I})) \tag{1}$$

Eq 1 was chosen such that: (1) if the CFG scales for image w_i and text prompt w_p are both equal to 1, the resulting vector $\tilde{\mathbf{X}}$ should be equal to the prediction $\mathbf{X}(\mathbf{I}, \mathbf{p})$ conditioned on the image and text, without Classifier Free Guidance. (2) if the CFG scales for image w_i and text w_p are both equal to 0, the resulting vector $\tilde{\mathbf{X}}$ should be equal to the un-conditioned prediction $\mathbf{X}(\varnothing)$.

In Eq 1 there is an ordering on the conditionings. We also considered alternate orderings in which we start with the text conditioning first instead of the image conditioning:

$$\tilde{\mathbf{X}} = \mathbf{X} + w_p(\mathbf{X}(\mathbf{p}) - \mathbf{X}(\varnothing)) + w_i(\mathbf{X}(\mathbf{I}, \mathbf{p}) - \mathbf{X}(\mathbf{p})) \tag{2}$$

Eq 2 did not lead to improvement over Eq 1, but required significantly different values for w_i and w_p to work equally well. We also considered formulas without ordering between the two conditionings, for instance:

$$\tilde{\mathbf{X}} = \mathbf{X} + w_i(\mathbf{X}(\mathbf{I}) - x(\varnothing)) + w_p(\mathbf{X}(\mathbf{p}) - x(\varnothing))$$

and

$$\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I}, \mathbf{p}) + w_i'(\mathbf{X}(\mathbf{I}, \mathbf{p}) - \mathbf{X}(\mathbf{p})) + w_p'(\mathbf{X}(\mathbf{I}, \mathbf{p}) - x(\mathbf{I}))$$

where $w_i' = (w_i - 1)$ and $w_p' = (w_p - 1)$

Similar to Eq 2, those formulas did not improve over Eq 1, and in addition miss the useful properties listed above.

Selecting CFG scales. Eq 1 requires to find the guidance factor w_i for image and w_p for text. We found that these factors influence the motion in the generated videos. To quantify this, we measure a 'motion score' on the generated videos by computing the mean energy of the motion vectors in the resulting `H.264` encoding. We found that the motion score was a good proxy for the amount of

Setting	Training stage			
	256px	512px HQ FT		FI
	$\mathcal{F}$	$\mathcal{F}$	$\mathcal{F}$	$\mathcal{I}$
Diffusion settings:				
Loss	Mean Squared Error			
Timesteps	1000			
Noise Schedule	quad	quad*		
Beta start	8.5×10^{-4}	$8.5 \times 10^{-4^*}$		
Beta end	1.2×10^{-2}	$1.2 \times 10^{-2^*}$		
Var type	Fixed small			
Prediction mode	eps-pred	v-pred		
0-term-SNR rescale	False	True [51]		
Optimizer	AdamW [52]			
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$			
Learning rate:				
Schedule	Constant			
Warmup Schedule	Linear			
Peak	1e-4		2.5e-5	1.5e-4
Warmup Steps	1K		10K	1.5K
Weight decay	0.0		1e-4	0.0
Dataset size	34M		1.6K	34M
Batch size	512		64	384
Transforms:				
Clip Sampler	Uniform			
Frame Sampler	Uniform			
Resize				
interpolation	Box + Bicubic			
size	256px	512px		
Center Crop	256px	512px		
Normalize Range	[-1, 1]			

Table 3: Training hyperparameters for various stages in our pipeline: 256px training, 512px training, High Quality finetuning (HQ FT), and frame interpolation (FI). *: noise schedules are changed afterwards with zero terminal-SNR rescaling [51].

motion, but did not provide signal into consistency of the motion. Higher motion as computed through motion vectors does not necessarily translate to interesting movement, as it could be undesirable jitter, or reflect poor object consistency. Table 4 shows how the CFG scales directly influence the amount of motion in the generated videos.

After narrowing down a few CFG value combinations by looking at the resulting motion score, we identified the best values by visual inspection and human studies. Qualitatively, we found that the (1) higher w_i for a fixed w_p, the more the model stays close to the initial image and favors camera motion; and (2) the higher w_p for a fixed w_i, the more the model favors movement at the expense of object consistency.

Model	w_p	w_i	Motion Score
w/o HQ finetuning	2.0	1.0	1.87
w/o HQ finetuning	8.0	1.0	2.87
w/o HQ finetuning	16.0	1.0	3.86
w/o HQ finetuning	8.0	1.0	2.87
w/o HQ finetuning	8.0	2.0	0.61
w/o HQ finetuning	8.0	3.0	0.25
HQ finetuned	2.0	2.0	11.1
HQ finetuned	8.0	2.0	12.7
HQ finetuned	16.0	2.0	13.5
HQ finetuned	8.0	1.0	14.9
HQ finetuned	8.0	2.0	12.7
HQ finetuned	8.0	3.0	11.3

Table 4: We measure the amount of motion in the generated videos using an automated motion score where a higher value reflects more motion. We use the prompts from [68]. The ratio of text CFG scale w_p to image CFG scale w_i influences the amount of motion in the video. We also observe that, w/o HQ fine-tuning, motion is much less and that the relative effect of CFG scales is even more pronounced.

Frame Interpolation Model. Here, we include extra details on the frame interpolation model, $\mathcal{I}$. First we explain our masked zero-interleaving strategy. Second we explain how we interpolate 16-frame 4fps videos from $\mathcal{F}$. § 3.3 in the main paper details how $\mathcal{I}$ is trained to take 8 zero-interleaved frames (generated from $\mathcal{F}$ at 4fps) as conditioning input and generate 37 frames at 16fps. One option for training an interpolation model that increases the fps by 4-fold is to generate 3 new frames between every pair of input frames (as in [6]). However, the downside to this approach is that the resulting interpolated video has a slightly shorter duration than the input video (since every input frame has 3 new generated frames after it, except the last input frame). We instead take the approach of using $\mathcal{I}$ to increase the duration of the input video, and we design a zero-interleaving scheme accordingly. Our interpolation model is trained to generate 3 new frames between every pair of frames, and also 4 new frames either side of the input video. As a result, during training $\mathcal{I}$ takes as conditioning input a 2s video, and generates a 2.3s video.

For interpolating 16-frame input videos from $\mathcal{F}$ (as described in § 4.2 in the main paper), we simply split the videos into two 8-frame videos and run interpolation on both independendly. In order to construct our final interpolated video, we discard the *overlapping* frames (the last 5 frames of the first interpolated video, and the first 4 of the second), and concatenate the two videos frame-wise. The resulting interpolated video is 65 frames long at 16fps (4.06 seconds in duration – we refer to these videos as 4 seconds long in the main paper for brevity).

2 Additional experiments

We detail additional experiments, viz. (i) an investigation into the effect of the initial image on our video generations, (ii) a quantitative comparison to prior work in image animation with automated metrics, (iii) a joint investigation into the effect of the number of training steps and data, and finally (iv) an analysis into the effect of the amount of training data.

Method	#Prompts	Q	F
Gen2 *vs*. Gen2 I2V	65 [6]	41.5	44.6
Emu Video *vs*. Gen2 I2V		72.3	78.4
Emu Video *vs*. Gen2		78.5	87.7

Table 5: Image conditioning for commercial T2V We compare Emu Video against two video generation variants of Gen2 API: (1) Gen2 which accepts only a text prompt as input and (2) Gen2 I2V which accepts an input image (generated using [57]) and a text prompt. We observe that the second variant (Gen2 I2V) outperforms the text-to-video Gen2 variant. Emu Video's generations are strongly preferred to both the variants of the Gen2 API.

Image conditioning for commercial T2V systems. We study the effect of image conditioning on the commercial T2V solution from Gen2 [54] in Table 5. The Gen2 API has two video generation variants: (1) A pure T2V API that accepts a text prompt as input and generates a video; and (2) an "image + text" API, denoted as Gen2 I2V, that takes an image and a text prompt as input to generate a video. We use images generated from [57] for the Gen2 I2V variant.

We observe that the Gen2 I2V variant outperforms the Gen2 API that only accepts a text prompt as input. We benchmark Emu Video against both variants of the API and observe that it outperforms Gen2 and the stronger Gen2 I2V API. In Table 3, we also compare Emu Video using the same images as Gen2 I2V for "image animation" and observe that Emu Video outperforms Gen2 I2V in that setting as well.

Automated metrics for image animation. We follow the setting from Table 3 and report automated metrics for comparison in Table 6. Following [22,78], we report Frame consistency (FC) and Text consistency (TC). We also report CLIP Image similarity [10] (IC) to measure the fidelity of generated frames to the conditioned image. We use CLIP ViT-B/32 model for all the metrics. Compared to VideoComposer [78], Emu Video generates smoother motion, as measure by frame consistency, maintains a higher faithfulness to the conditioned image, as measured by the image score, while adhering to the text on both the prompt sets. Emu Video fares slightly lower compared to PikaLabs and Gen2 on all three metrics. Upon further inspection, Emu Video (motion score of 4.98) generates more motion compared to PikaLabs and Gen2 (motion scores of 0.63 and 3.29 respectively). Frame and image consistency favour static videos resulting in the lower scores of Emu Video on these metrics.

Method	Dataset	FC (↑)	IC (↑)	TC(↑)
VideoComposer [78]	AYL [6]	96.8	86.4	33.3
PikaLabs I2V		99.9	95.0	34.6
Gen2 I2V		99.9	96.8	34.3
Emu Video		99.3	94.2	34.2
VideoComposer [78]	MAV [68]	95.2	82.6	31.3
Emu Video		98.9	91.3	32.1

Table 6: Automatic evaluation of Emu Video *vs.* prior work in text-conditioned image animation. We compare Emu Video against three contemporary methods following the settings from 3 using Frame consistency (FC), Image similarity (IC), and Text consistency (TC). Emu Video outperforms VideoComposer across both the prompt sets and all three metrics. Automatic metrics favor static videos to ones with motion, resulting in lower scores for Emu Video compared to PikaLabs and Gen2.

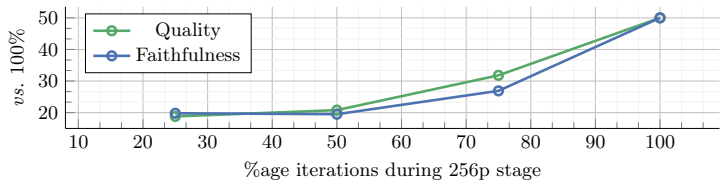


Fig. 1: Performance *vs.* training iterations. On training the 256px stage for fewer iterations, we compare the generations after the same 512px finetuning to the 100% trained model via human evaluations. We observe a gradual drop in performance, indicating the importance of the low-resolution high-FPS pretraining stage.

Effect of the number of training steps and data. In Figure 1, we vary the number of training steps in the initial low-resolution high-FPS pretraining stage. Note that since we run one full epoch through the data during this training stage, reducing the steps correspondingly also reduces the amount of training data seen. We finetune each of these models at higher resolution/low FPS (512px, 4fps) for the same (small) number of steps – 15K. We compare the model trained with 100% low-resolution pretraining with models with less low-resolution pretraining using human evaluations. We observe a gradual drop in performance as we reduce the low-resolution pretraining iterations to 75%, 50% and 25%, indicating the importance of that stage.

Effect of the amount of training data. In Figure 2, we vary the amount of training data, while keeping the training iterations fixed for both the training stages, and perform a similar comparison as in Figure 1. Here we find a much smaller drop in performance as we reduce the amount of data. This suggests that Emu Video can be trained effectively with relatively much smaller datasets, as long as the model is trained long enough (in terms of training steps).

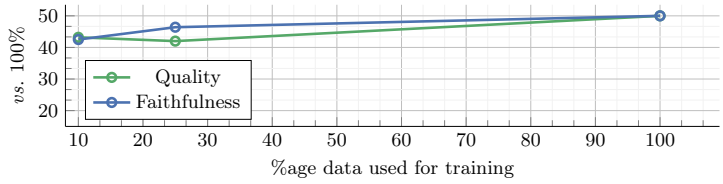


Fig. 2: Performance *vs.* training data. We train our model with less data (for both 256px and 512px stages) while keeping the training steps constant, and compare the generations with the the 100% data model via human evaluations. We observe that even with 10% data, we only see a slight degradation in performance ($\sim$ 43% on both Quality and Faithfulness), showcasing that our method works well even with a fraction of the data.

Source	#prompts
Make-A-Video [68]	307
Imagen Video [35]	55
Align Your Latents [6]	65
PYOCO [30]	74
Reuse & Diffuse [31]	23

Table 7: Text prompt sets used for evaluation in our work. We use the text prompt sets from prior work to generate videos.

3 Human evaluations

We rely on human evaluations for making quantitative comparisons to prior work. In Sec. 4 in the main paper, we introduce our method for robust human evaluations. We now give extra details on this method, termed JUICE, and analyse how it improves robustness, and explain how we ensure fairness in the evaluations. Additionally, in Table 7 we summarize the prompt datasets used for evaluations.

3.1 Robust Human Evaluations with JUICE

When comparing to prior work, we use human evaluations to compare the generations from pairs of models. Unlike the naive approach, where evaluators simply pick their choice from a pair of generations, we ask the evaluators to select a reason when making their choice. We call this approach JUICE, where evaluators are asked to 'justify your choice'. We show an example of the templates used for human evaluations for both video quality and text faithfulness in Figure 3, where the different possible justifying reasons are shown. One challenge faced when asking evaluators to justify their choice is that human evaluators who are not experts in video generation may not understand what is meant by terms such as "Object/scene consistency" or "Temporal text alignment" or may have subjective interpretations, which would reduce the robustness of the evaluations. To alleviate this challenge, for each justifying option we show the human evaluators examples of generated video comparisons where each of the

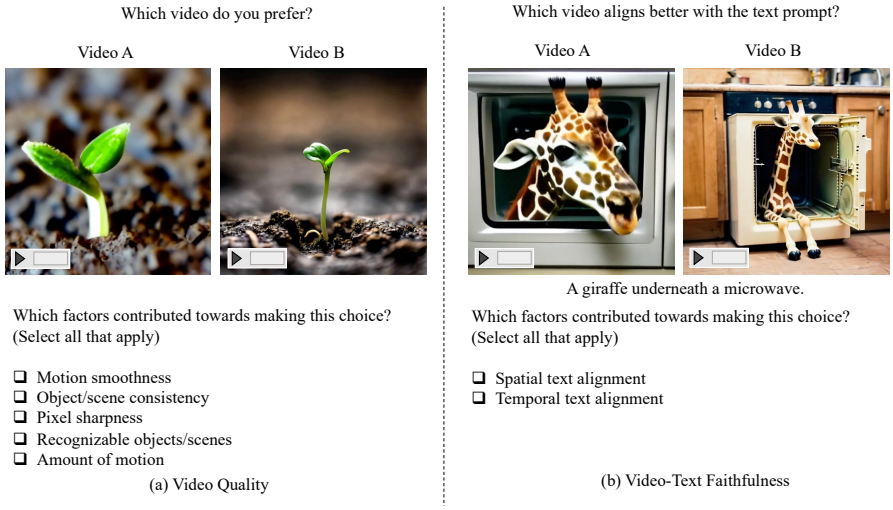


Fig. 3: The JUICE template to compare two models in terms of (a) video quality and (b) video-text alignment. Here, human evaluators must justify their choice of which generated video is superior through the selection of one or more contributing factors, shown here. To ensure that human evaluators have the same understanding of what these factors mean, we additionally provide training examples of video comparisons where each of the justifying factors could be used in selecting a winner.

factors could be used is used in determining a winner. It is important that when giving human evaluators training examples such as these that we do not bias them towards Emu Video's generations over those of prior work. Thus, to ensure fairness in the comparisons, we make sure that these training examples include cases where generated videos from different prior works are superior to Emu Video and vice-versa. As detailed in the main paper, for each comparison between two videos from two different models, we use the majority vote from 5 different human evaluators. To further reduce annotator bias we make sure that the relative positioning of the generated videos being shown to the human evaluators is randomized. For details on how we ensure fairness in human evaluations when comparing videos with different resolutions, see Sec. 4.

Next, we analyze quantitatively how JUICE improves human evaluation reliability and robustness. To identify unbiased JUICE factors differentiating any two video generation models on Quality and Faithfulness, we made an initial pool of random video samples generated by a few models, and asked internal human raters to explicitly explain their reasoning for picking one model over another. We then categorized them into five reasons for Quality and two for Faithfulness as mentioned in Section 3.2.

Effect of JUICE on improving evaluation reliability and robustness of human evaluations. We measure the reliability of our human evaluations when evaluators are required to justify their choice. For each pair of videos

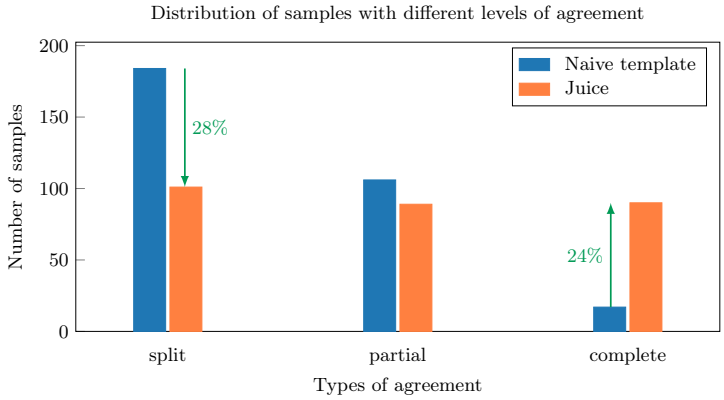


Fig. 4: Human agreement in Emu Video *vs.* Make-A-Video. Distribution of samples with 'split' (2|3 or 3|2 votes), 'partial' (4|1 or 1|4 votes), or 'complete' (5|0 or 0|5 votes) agreement when using a naive evaluation *vs.* JUICE. Our JUICE evaluation reduces ambiguity in the task and results in a 28% reduction in the number of samples with 'split' agreement and a 24% increase in the number of samples with 'complete' agreement. This improves Fleiss' kappa from 0.004 to 0.31.

which are compared, we look at the votes for model A *vs.* model B and call the agreement between annotators either 'split' (2|3 or 3|2 votes), 'partial' (4|1 or 1|4 votes), or 'complete' (5|0 or 0|5 votes). We run human evaluations comparing our generations *vs.* Make-A-Video, first using a naive evaluation template and then with JUICE, and show the results in Figure 4. We observe that the number of samples with 'split' agreement is decreased significantly by 28%, and the number of 'complete' agreements is increased by 24%.

Next, we use Fleiss' kappa [26] as a statistical measure for inter-rater reliability for a fixed number of raters. This metric stands for the amount by which the observed agreement exceeds the agreement by chance, *i.e.*, when the evaluators made their choices completely randomly. Fleiss' kappa works for any number of evaluators giving categorical ratings and we show the values in Figure 5. The value of kappa is always in the range of $[-1, 1]$, with positive kappa values representing an agreement. To better understand its behavior and range of scores in our evaluation setup, we perform an experiment on a simulated data representing our specific case of 304 tasks with two classes, model A-vs-B, and five evaluators per task. We begin with computing the kappa value when we have a 'complete' agreement among evaluators on all tasks, i.e. when all five evaluators choose either model A or model B in each task. This run receives a kappa value of 1 (blue dot in Figure 5). We gradually decrease the number of samples with complete agreement by introducing samples with 'partial' agreement when four out of five evaluators picked model A or model B (green line in Figure 5) Similarly, we decrease the number of samples with complete agreement by replacing them with samples where three out of the five evaluators picked model A or model B, illustrated with a red line. As shown in the plot, the kappa value ranges

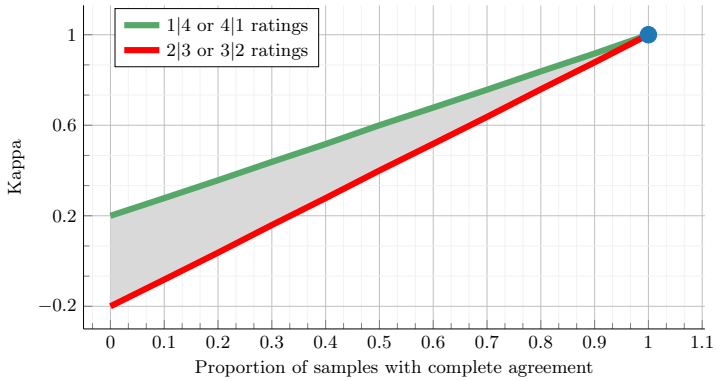
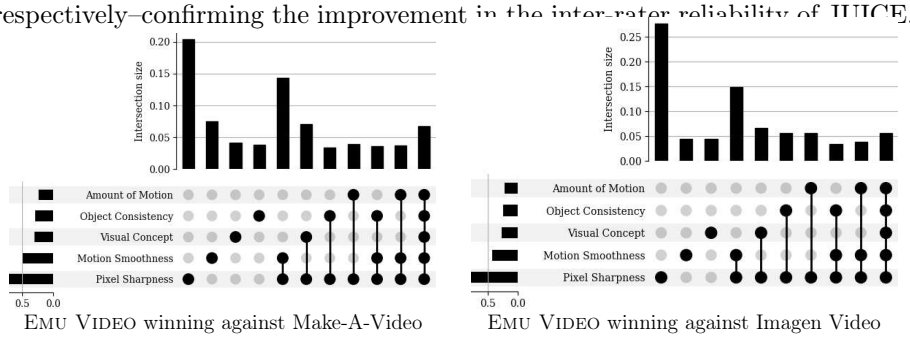


Fig. 5: Analysis of Fleiss' kappa for a simulated two-class five-raters evaluation task. The blue dot shows the kappa value when we have a complete agreement among evaluators on all the samples. We progressively replace samples with 5|0 or 0|5 votes (complete agreeement) with either 1|4 or 4|1 or 3|2 or 2|3 votes and compute the Fleiss' kappa (shown in green and red). The shaded region shows the kappa value for different proportions of samples with complete, partial or split agreements.

from −0.2 (ratings always being 'split') to 1.0 (ratings always having 'complete' agreement). Different proportions of samples with 'complete', 'partial' or 'split' agreements result in a kappa value in the shaded region. We compute and compare kappa values for the naive evaluation and JUICE evaluation–0.004 and 0.31, respectively–confirming the improvement in the inter-rater reliability of JUICE.



Emu Video winning against Make-A-Video

Emu Video winning against Imagen Video

Fig. 6: Vertical bars show percentage of each reason and its co-occurrence with other reasons picked for Emu Video against Make-A-Video (left) and Imagen Video (right). Horizontal bars depict the overall percentage of each reason, similar to Figure 6. Pixel sharpness and motion smoothness are the two most contributing factors in the Emu Video win against both baselines.

Analyzing human evaluations. To clearly understand the strengths of each model in our evaluations, we find the most contributing factors when Emu Video generations are preferred to each baseline in Figures 6, 7. A more detailed dis-

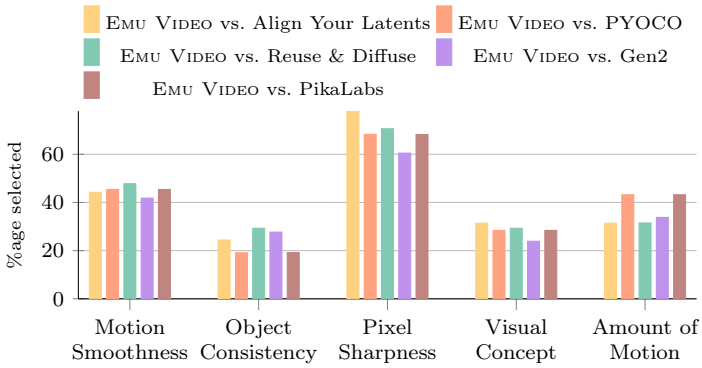


Fig. 7: Percentage of each reason selected for samples where Emu Video wins against each baseline model on Quality. Reasons that human evaluators pick Emu Video generations over the baseline models from Figure 2 are primarily pixel sharpness and motion smoothness of our videos for most models. Amount of motion in Emu Video generations is also an impactful winning factor against PYOCO and PikaLabs.

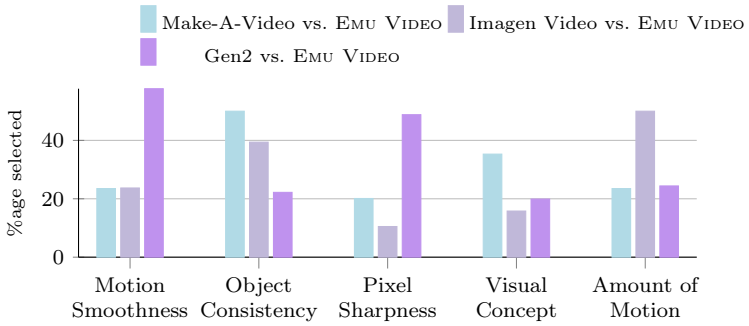


Fig. 8: Percentage of each reason selected for samples where each baseline model wins against Emu Video on Quality. Among the few preferred Make-A-Video generations from Figure 2 against Emu Video, object consistency has been the primary reason, while for Imagen Video generations, amount of motion has been an additional considerable reason. Gen2 generations preferred over Emu Video are mainly selected due to their motion smoothness and pixel sharpness.

tribution of each reason and its co-occurrence with other factors is illustrated in Figure 6. We similarly, plot the percentage of each reason picked for the best three baseline generations preferred to Emu Video in Figure 8.

4 Comparisons to Prior Work

In § 4.2 in the main paper, we conduct human evaluations comparing Emu Video to prior work. Here, we share further details and include human evaluation results using a different setup. Specifically, in § 4.1 we outline the prompt datasets that are used in comparisons to prior work. In § 4.2 we detail how we sampled from the commercial models that we compare to in the main paper. In § 4.3 we give details on the postprocessing done for the human evaluations in Figure 2 in the main paper. In § 4.4 we include further human evaluations conducted without postprocessing the videos from Emu Video or prior work.

4.1 Datasets used for Prior Work Comparisons

Since many of the methods that we compare to in Figure 2 are closed source, we cannot generate samples from all of them with one unified prompt dataset, and instead must construct different datasets via each method's repsective publicly released example generated videos. In total, we use 5 different prompt datasets. The human evaluations in Figure 2 for Make-A-Video, Imagen Video, Align Your Latents, PYOCO, and Reuse & Diffuse were conducted using the prompt datasets from the respective papers (see Table 7 for details). Certain methods that we compare to are either open-source (CogVideo) or can be sampled from through an online interface (Gen2 and Pika Labs). For these, human evaluations are conducted using the prompt set from Align Your Latents.

4.2 Sampling from Commercial Models

The commercially engineered black-box text-to-video models that we compare to (Pika Labs and Gen2) can be sampled from through an online interface. Here we include details for how we sampled from these models. In both cases, these interfaces allow for certain hyper-parameters to be chosen which guide the generations.

We selected optimal parameters for each of the models by varying the parameters over multiple generations and choosing those that consistently resulted in the best generations. For Pika Labs, we use the arguments "-ar 1:1 -motion 2" for specifying the aspect ratio and motion. For Gen2, we use the "interpolate" and "upscale" arguments and a "General Motion" score of 5. All samples were generated on October 24th 2023.

4.3 Postprocessing Videos for Comparison

Our goal with our main human evaluations in Figure 2 is to ensure fairness and reduce any human evaluator bias. To ensure this fairness, we postprocess the

Model	Video Dimensions		
	T × H × W	Frame Rate	Duration (s)
Emu Video	65 × 512 × 512	16	4.06
Pika	72 × 768 × 768	24	3.00
Gen2	96 × 1024 × 1792	24	4.00
CogVideo	32 × 480 × 480	8	4.00
Reuse & Diffuse	29 × 512 × 512	24	1.21
PYOCO	76 × 1024 × 1024	16	4.75
Align Your Latents	112 × 1280 × 2048	30	3.73
Imagen Video	128 × 768 × 1280	24	5.33
Make-A-Video	92 × 1024 × 1024	24	3.83
VideoComposer	16 × 256 × 256	8	2

Table 8: Video Dimensions. The dimensions of the generated videos from Emu Video and each of the prior work. The top and bottom part of the table shows the specifications of Text-to-Video and Image-to-Video models respectively. Each of the prior works generates videos at different dimensions, making unbiased human evaluation a challenge.

Models Compared	Dimensions after Postprocessing		
	T × H × W	Frame Rate	Duration (s)
Emu Video vs. Pika Labs	48 × 512 × 512	16	3.00
Emu Video vs. Gen2	65 × 512 × 512	16	4.06
Emu Video vs. CogVideo	32 × 480 × 480	8	4.00
Emu Video vs. Reuse & Diffuse	19 × 512 × 512	16	1.19
Emu Video vs. PYOCO	65 × 512 × 512	16	4.06
Emu Video vs. Align Your Latents	65 × 512 × 512	16	4.06
Emu Video vs. Imagen Video	65 × 512 × 512	16	4.06
Emu Video vs. Make-A-Video	61 × 512 × 512	16	3.81
Emu Video vs. VideoComposer	16 × 256 × 256	8	2

Table 9: Video Dimensions after postprocessing for human evaluations.. To ensure fairness in the human evaluations in in Figure 2 in the main paper, we postprocess the videos for each comparison so that they have equal dimensions and hence are indistinguishable aside from their generated content. The top and bottom part of the table shows the specifications of Text-to-Video and Image-to-Video models respectively.

videos from each model being compared (as outlined in § 4.2 in the main paper). Here, we give further details on the motivation behind this decision, and explain how this postprocessing is done. Results for human evaluations conducted without any postprocessing are discussed in § 4.4.

As outlined in Sec. 3, our human evaluations are conducted by showing evaluators repeated comparisons of videos generated by two different models for the same prompt, and asking them which model they prefer in terms of the metric be-

	Make-A-Video	Imagen Video	Align Your Latents	PYOCO	Reuse & Diffuse	CogVideo	Gen2	PikaLabs
#Prompts	307 [68]	55 [35]	65 [6]	74 [30]	23 [31]	65 [6]	65 [6]	65 [6]
Quality	96.8	90.9	96.9	93.2	95.7	100.0	83.1	93.9
Faithfulness	86.0	69.1	90.8	89.2	100.0	100.0	98.5	100.0

Table 10: Emu Video *vs*. prior work where videos are not postprocessed. We evaluate text-to-video generation in terms of video quality and text faithfulness win-rates evaluated by the majority votes of human evaluators for Emu Video vs. Prior work methods. We compare methods here with their original dimensions (aspect ratio, duration, frame rate). Emu Video significantly outperforms all prior work across all settings and metrics.

ing evaluated. It is key for the fairness of the human evaluation that the evaluator treats each comparison independently. It is hence important that the evaluator does not know which model generated which video, otherwise they can become biased towards one model over the other. Since each method generates videos at different dimensions (see Table 8), conducting the human evaluations without postprocessing the videos would lead to this annotator bias. Hence we decide to postprocess the videos being compared such that they have the same aspect-ratios, dimensions and frame rates so that they are indistinguishable aside from their generated content. For each pair of models being compared, we downsample these dimensions to the minimum value between the two models (see Table 9 for details). Next, we detail how we postprocess the videos.

Aspect Ratio. Since Emu Video generates videos at a 1:1 aspect ratio, all videos are postprocessed to a 1:1 aspect ratio by centre cropping.

Spatial Dimension. The height and width of videos are adjusted using bilinear interpolation.

Video Duration. The duration of videos is reduced via temporal centre cropping.

Frame rate. The frame rate is adjusted using torchvision. The number of frames is selected according to the desired frame rate and video duration.

Next we discuss human evaluation results where videos are compared without any postprocessing.

4.4 Prior Work at Original Dimensions

In this Section, we include further human evaluation results between Emu Video and prior work where we do not perform any postprocessing on the videos and conduct the evaluations with the original dimensions (as detailed in Table 8). In this system-level comparison, human evaluators are comparing between videos that may have very different aspect ratios, durations, and frame rates, and in turn may become biased towards one model over another after seeing repeated comparisons. We note that since the dimensions of the videos here are so large, we must scale the height of each video so that both compared videos can fit on one screen for human evaluators. All other dimensions remain as in the original sampled videos. The results are in Table 10. Similar to the human evaluations conducted with postprocessed videos in Figure 2 in the main paper,

Emu Video significantly outperforms prior work in terms of both text faithfulness and video quality. Even when comparing Emu Video's generated videos to generated videos with longer durations (including PYOCO, Imagen Video), wider aspect ratios (incliding Gen2, Align Your Latents), or higher frame rates (including Pika, Gen2), human evaluators still prefer Emu Video's generated videos in both metrics. We hypothesize that the vastly improved frame quality and temporal consistency of Emu Video still outweighs any benefits that come from any larger dimensions in the prior work's videos.

Interestingly, Emu Video wins by larger margins here than in the postprocessed setting (an average win rate of 93.8% in quality and 93.1% in faithfulness here, vs. 91.8% and 86.6% in the postprocessed comparison). We conjecture that this improvement in win rates for Emu Video may be due to the potential evaluator bias introduced in this evaluation setting. This introduced bias tends to favor Emu Video since our video generations are on average superior in terms of quality and faithfulness than those of prior work. Hence in this paper we primarily report and refer to the human evaluation scores from the fairer postprocessed setting.

5 Qualitative Results

In this Section, we include additional qualitative results from Emu Video (in § 5.1), and further qualitative comparisons between Emu Video and prior work (in § 5.2)

5.1 Further Emu Video qualitative Results

Examples of Emu Video's T2V generations are shown in Figure 9, and Emu Video's I2V generations are shown in Figure 10. As shown, Emu Video generates high quality video generations that are faithful to the text in T2V and to both the image and the text in I2V. The videos have high pixel sharpness, motion smoothness and object consistency, and are visually compelling. Emu Video generates high quality videos for both natural prompts and fantastical prompts. We hypothesize that this is because Emu Video is effectively able to retain the wide range of styles and diversity of the T2I model due to the factorized approach.

5.2 Qualitative Comparisons to Prior Work

We include further qualitative comparisons to prior work in Figs. 11, 12, 13, 14, 15 and 16. This Section complements § 4.2 in the main paper where we quantatively demonstrate via human evaluation that Emu Video significantly outperforms the prior work in both video quality and text faithfulness. Emu Video consistently generates videos that are significantly more text faithful (see Figs. 12 and 14), with greater motion smoothness and consistency (see Figs. 13 and 15), far higher pixel sharpess (see Figure 16), and that are overall more visually compelling (see Figure 11) than the prior work.

(Ours - Emu Video) *Prompt:* A hamster wearing virtual reality headsets is a dj in a disco.

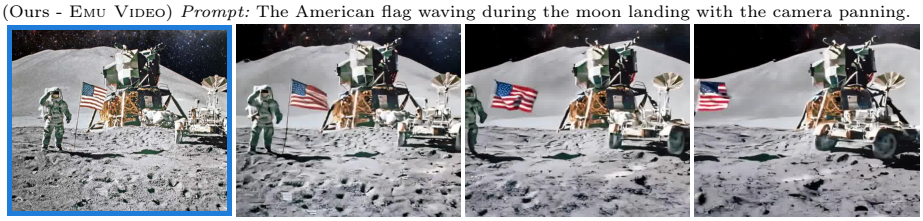
(Emu Video) *Prompt:* A massive tidal wave crashes dramatically against a rugged coastline.

(Emu Video) *Prompt:* A majestic white unicorn with a golden horn walking in slow-motion under water.

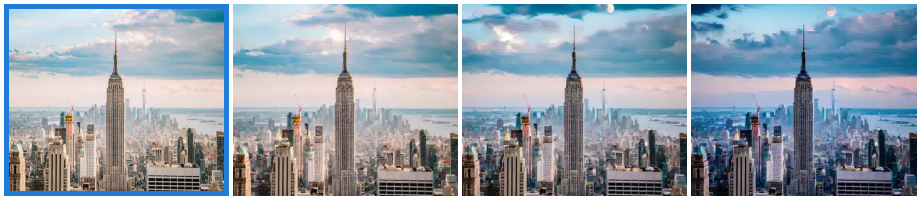
(Emu Video) *Prompt:* A grizzly bear hunting for fish in a river at the edge of a waterfall, photorealistic.

Fig. 9: Example T2V generations from Emu Video for a selection of diverse prompts (shown above each row of frames). Emu Video generates natural-looking videos which are faithful to the text and high in visual quality. The videos are highly temporally consistent, with smooth motion. Emu Video is able to generate high quality videos for both natural prompts (rows 2 and 4) depicting scenes from the natural world, and also fantasical prompts including DJing hamsters (row 1) and underwater unicorns (row 3).

(Ours - Emu Video) *Prompt:* The American flag waving during the moon landing with the camera panning.



(Emu Video) *Prompt:* The sun sets and the moon rises.

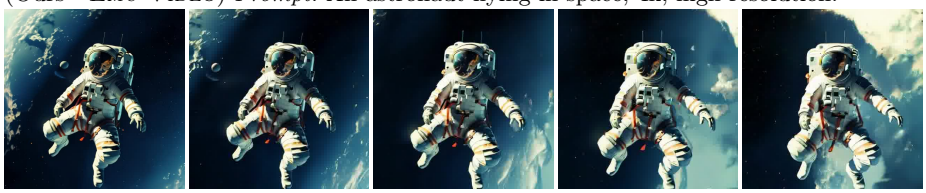


(Emu Video) *Prompt:* Satellite flies across the globe.

(Emu Video) *Prompt:* horse moving its legs.

Fig. 10: Example I2V generations from Emu Video for a selection of diverse prompts (shown above each row of frames). Emu Video generates natural-loooking videos from the conditioning image (shown in a blue box on the left side of each row of frames) and the text prompt, that have smooth and consistent motion.

(Ours - Emu Video) *Prompt:* An astronaut flying in space, 4k, high resolution.



(Gen2) *Prompt:* An astronaut flying in space, 4k, high resolution.

(PikaLabs) *Prompt:* An astronaut flying in space, 4k, high resolution.

(Align Your Latents) *Prompt:* An astronaut flying in space, 4k, high resolution.

(CogVideo) *Prompt:* An astronaut flying in space, 4k, high resolution.

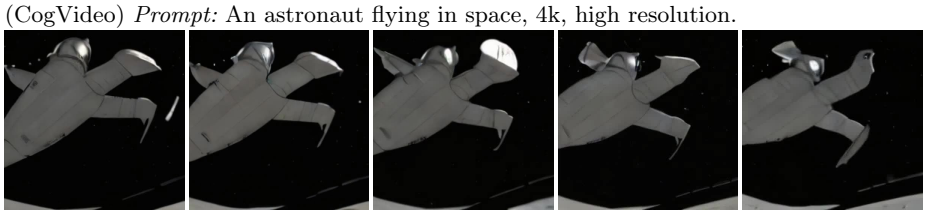


Fig. 11: Example T2V generations from Emu Video and a selection of prior work methods that we compare to in the main paper for the same prompt, namely Gen2, Pika Labs, Align your latents, and CogVideo. Emu Video generates higher quality videos that are more faithful to the text, have realistic & smooth movement, and are visually compelling. In this example, CogVideo cannot generate a natural-looking video (see 5th row). PikaLabs is not faithful to the text and does not generate a realistic looking astronaut (see 3rd row), whereas Align Your Latents generates a video with low visual quality. Gen2's video, although visually superior to other prior work, lacks pixel sharpness and is not as visually compelling as Emu Video.

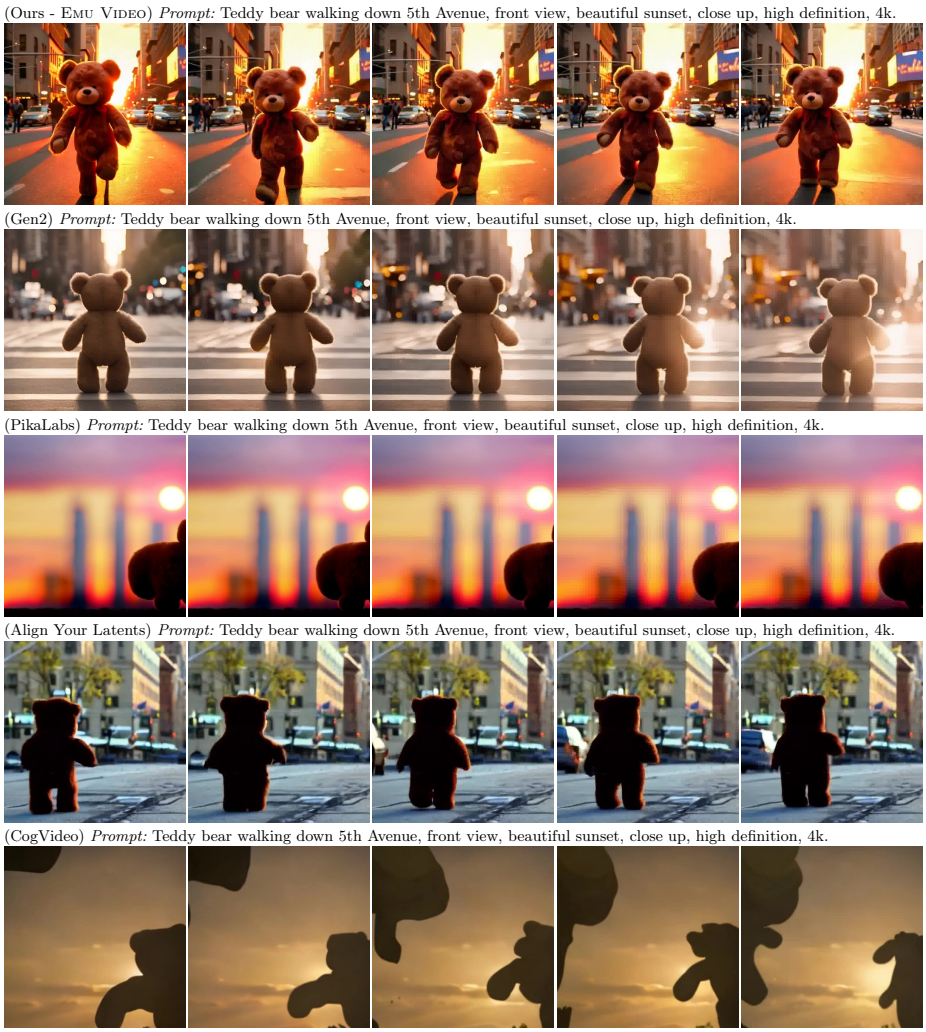


Fig. 12: Example T2V generations from Emu Video and a selection of prior work methods that we compare to in the main paper for the same prompt, namely Gen2, Pika Labs, Align your latents, and CogVideo. CogVideo and PikaLabs's videos are not faithful to the text and lack on visual quality. Gen2 correctly generates a video of a bear on a street, but the bear is not moving, and there is limited motion in the vidoeo. Align Your Latents's video lacks motion smoothness and pixel sharpness. On the other had, Emu Video's video has very high visual quality and high text faithfulness, with smooth and consistent high motion.

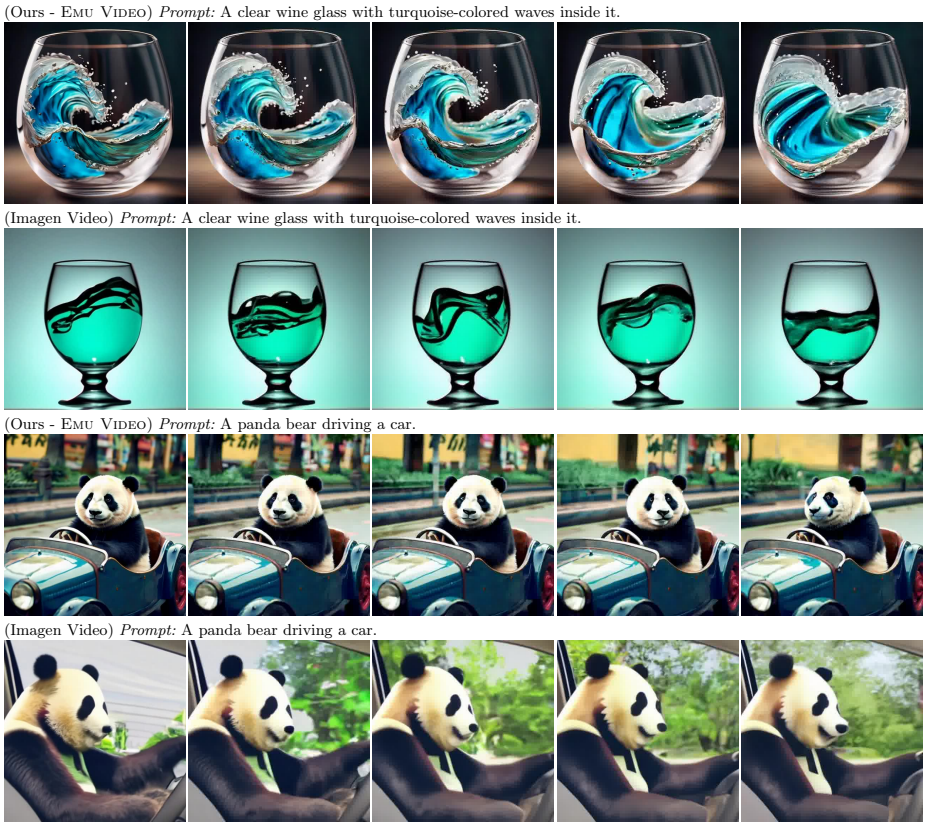
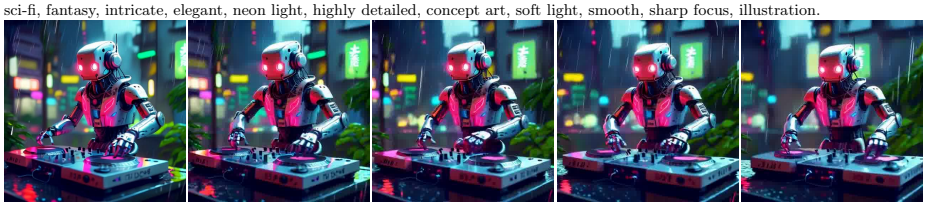


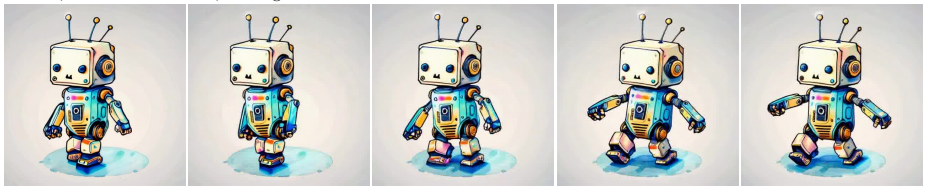
Fig. 13: Example T2V generations from Emu Video and Imagen Video on two prompts (which are shown above each row of frames). Imagen Video generates videos that are faithful to the text, however the videos lack in pixel sharpness and motion smoothness. Additionally Imagen Video's generations lack fine-grained high-quality details such as in the panda's hair (see 4th row) and the water movements (see 2nd row). Emu Video on the other hand generates high quality videos that are faithful to the text, and with high pixel sharpness and motion smoothness. Emu Video accurately generates natural looking fine-grained details such as the hair on the panda (see 3rd row) and the water droplets in the waves (see 1st row).

(Ours - Emu Video) *Prompt:* A robot dj is playing the turntable, in heavy raining futuristic tokyo rooftop cyberpunk night, sci-fi, fantasy, intricate, elegant, neon light, highly detailed, concept art, soft light, smooth, sharp focus, illustration.



(PYOCO) *Prompt:* A robot dj is playing the turntable, in heavy raining futuristic tokyo rooftop cyberpunk night, sci-fi, fantasy, intricate, elegant, neon light, highly detailed, concept art, soft light, smooth, sharp focus, illustration.

(Ours - Emu Video) *Prompt:* A cute funny robot dancing, centered, award winning watercolor pen illustration, detailed, isometric illustration, drawing.



(PYOCO) *Prompt:* A cute funny robot dancing, centered, award winning watercolor pen illustration, detailed, isometric illustration, drawing.

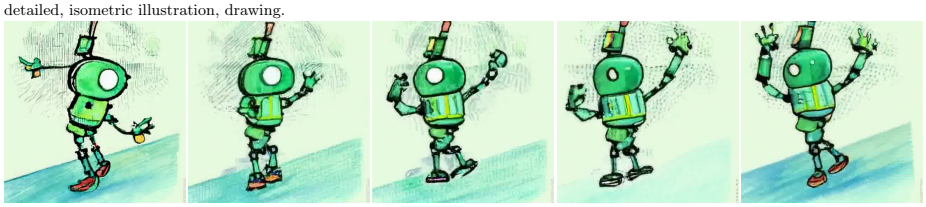


Fig. 14: Example T2V generations from Emu Video and PYOCO on two prompts (which are shown above each row of frames). Whereas PYOCO's videos lack motion smoothness or consistency and cannot generate fine-grained details, Emu Video instead generates highly realistic videos that are smooth and consistent. Emu Video can generate high quality videos given fantastical prompts.

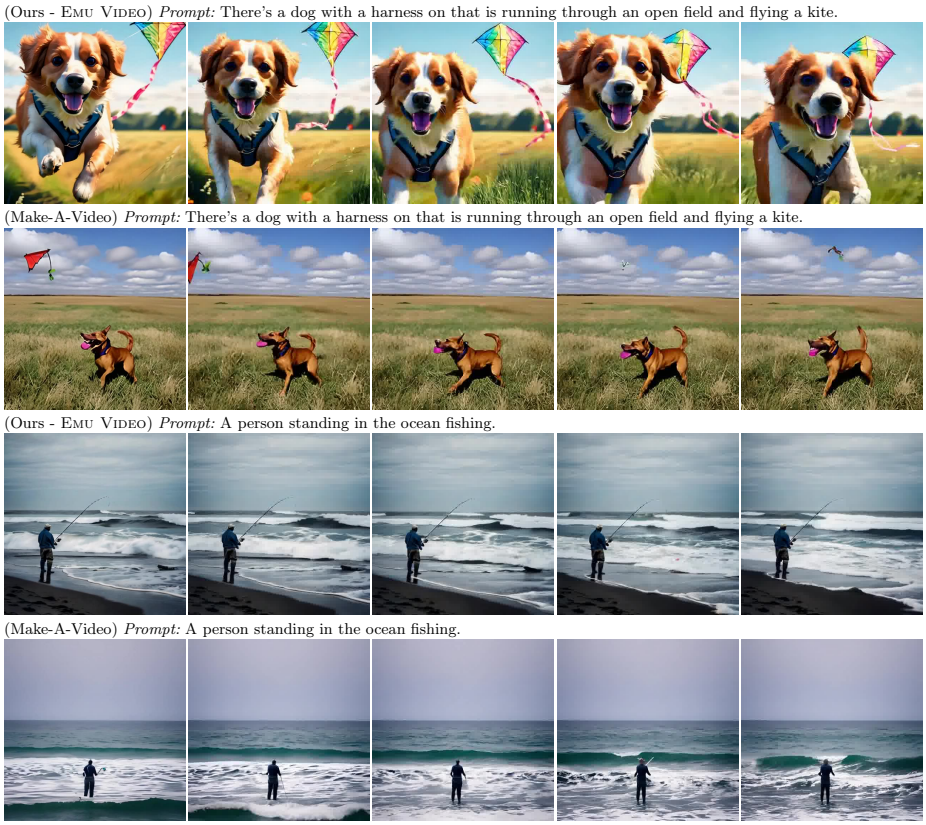


Fig. 15: Example T2V generations from Emu Video and Make-A-Video on two prompts (which are shown above each row of frames). whereas Make-A-Video's videos lack pixel sharpness and object consistency, Emu Video generates high quality and natural-looking videos. Emu Video's videos have high motion smoothness and object consistency.

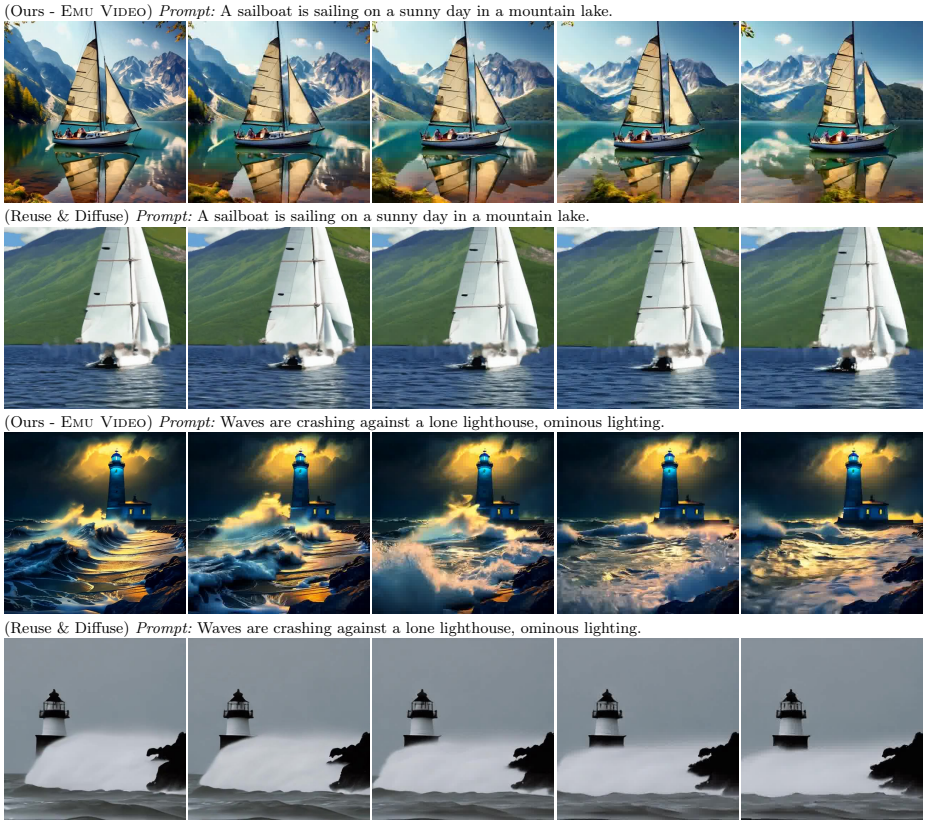


Fig. 16: Example T2V generations from Emu Video and Reuse & Diffuse on two prompts (which are shown above each row of frames). whereas Reuse & Diffuse's videos lack in visual quality both in terms of pixel sharpness, and temporal consistency, Emu Video instead generates visually compelling and natural-looking videos which accurately follow the prompt.

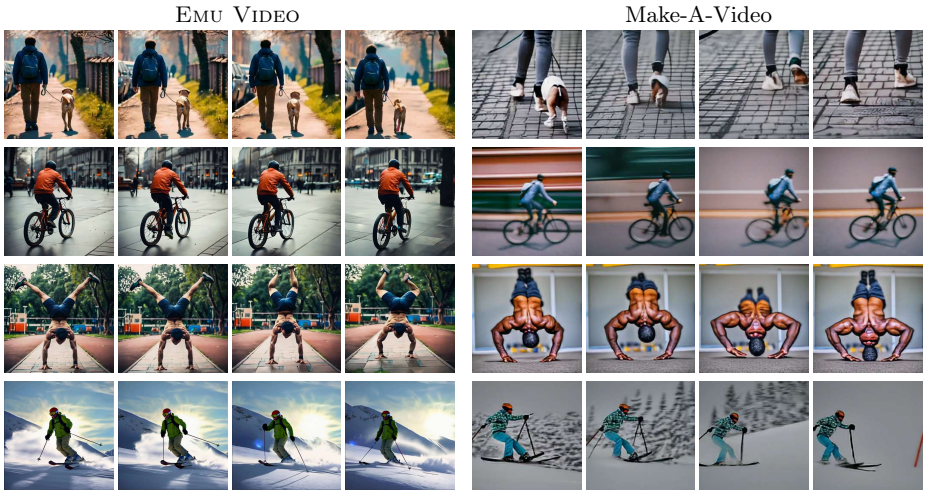


Fig. 17: Zero-Shot text-to-video generation on UCF101. The classes for these videos from top to bottom are: walking with a dog, biking, handstand pushups, skiing. Our generations are of higher quality and more coherent than those from Make-A-Video.

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)