

Lumiere: A Space-Time Diffusion Model for Video Generation

Omer Bar-Tal^{*1 2} Hila Chefer^{*1 3} Omer Tov^{*1} Charles Herrmann^{†1} Roni Paiss^{†1} Shiran Zada^{†1}
Ariel Ephrat^{†1} Junhwa Hur^{†1} Yuanzhen Li¹ Tomer Michaeli^{1 4} Oliver Wang¹ Deqing Sun¹ Tali Dekel^{1 2}
Inbar Mosseri^{†1}

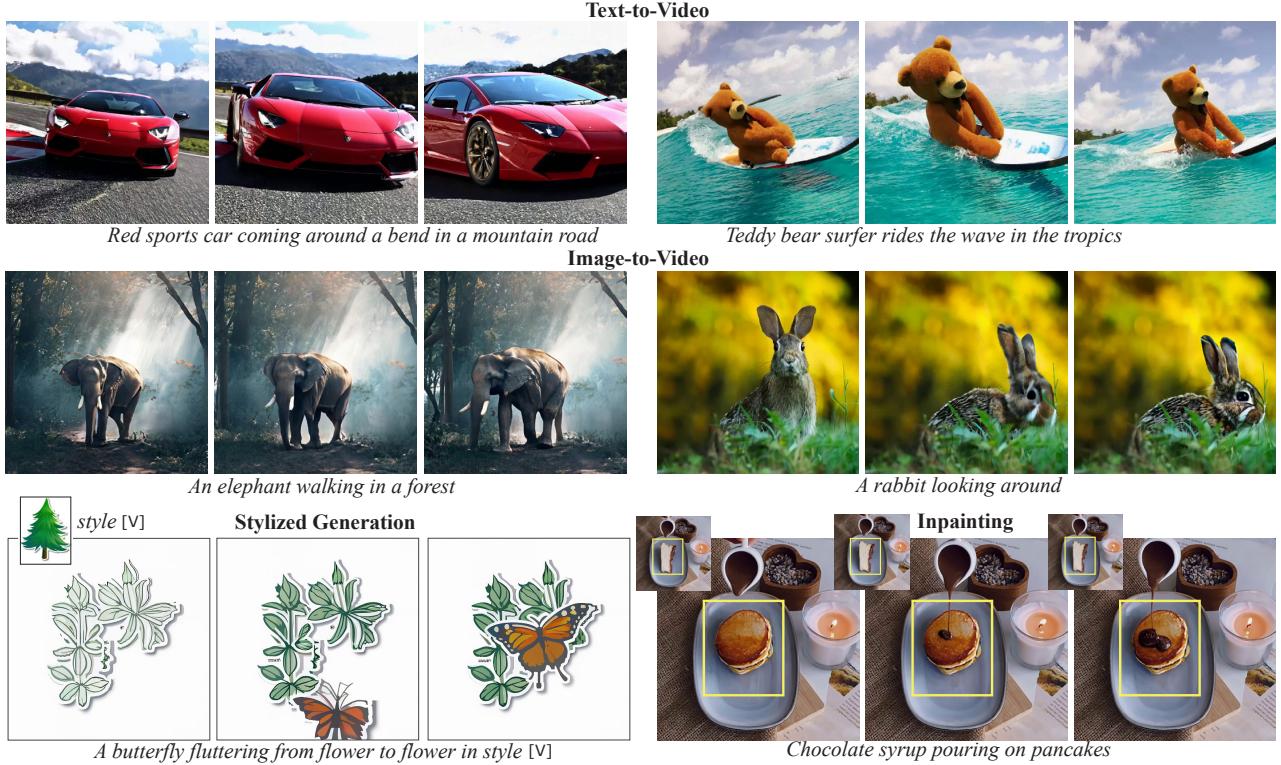
¹Google Research²Weizmann Institute³Tel-Aviv University⁴Technion

Figure 1: Sample results generated by *Lumiere*, including text-to-video generation (first row), image-to-video (second row), style-referenced generation, and video inpainting (third row; the bounding box indicates the inpainting mask region).

Abstract

We introduce *Lumiere* – a text-to-video diffusion model designed for synthesizing videos that portray realistic, diverse and coherent motion – a pivotal challenge in video synthesis. To this end, we introduce a Space-Time U-Net architecture that generates the entire temporal duration of the video *at once*, through a single pass in the model. This is in contrast to existing video models which synthesize distant keyframes followed by tem-

poral super-resolution – an approach that inherently makes global temporal consistency difficult to achieve. By deploying both spatial and (importantly) temporal down- and up-sampling and leveraging a pre-trained text-to-image diffusion model, our model learns to directly generate a full-frame-rate, low-resolution video by processing it in multiple space-time scales. We demonstrate state-of-the-art text-to-video generation results, and show that our design easily facilitates a wide range of content creation tasks and video editing applications, including image-to-video, video inpainting, and stylized generation.

^{*}Equal first author [†]Core technical contribution

Work was done while O. Bar-Tal, H. Chefer were interns at Google.
Webpage: <https://lumiere-video.github.io/>

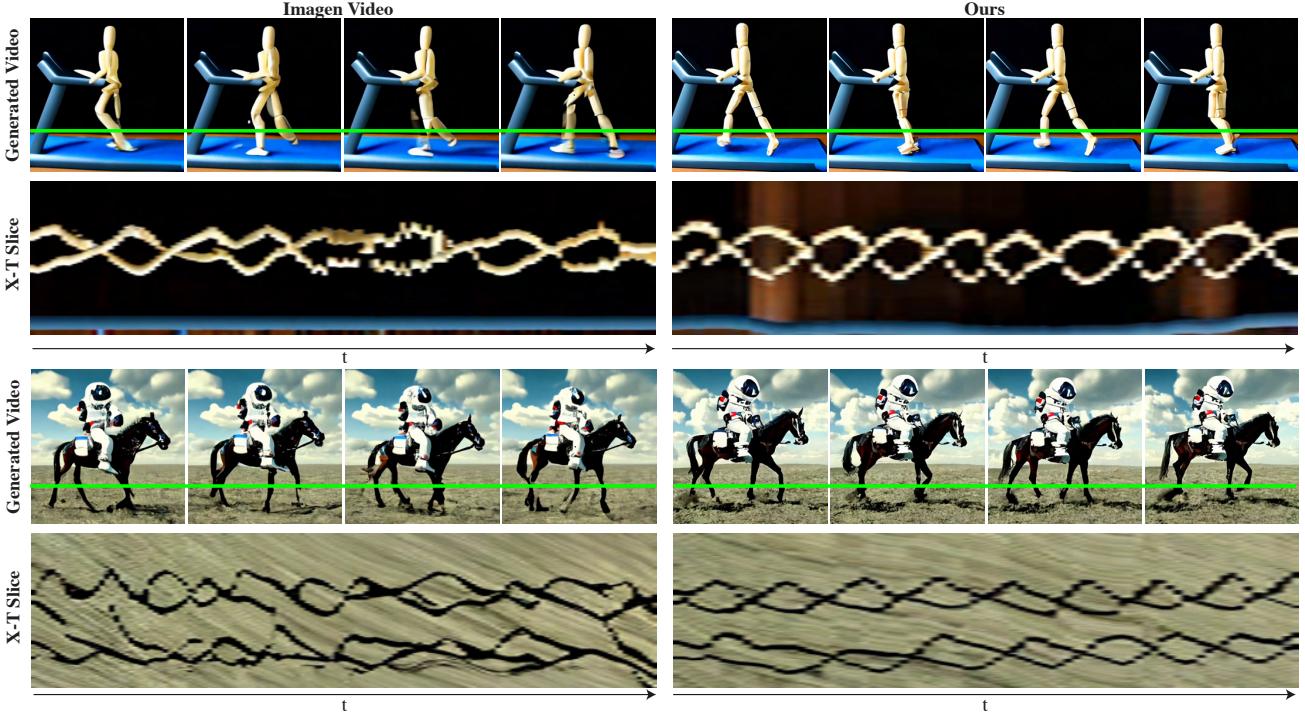


Figure 2: **Temporal consistency in generated videos.** Representative examples of generated videos using our model and ImagenVideo (Ho et al., 2022a) for periodic motion. We apply Lumiere image-to-video generation, conditioned on the first frame of a video generated by ImagenVideo, and visualize corresponding X-T slices. ImagenVideo struggles to generate globally coherent repetitive motion due to its cascaded design and temporal super resolution modules, which fail to resolve aliasing ambiguities consistently across temporal windows.

1. Introduction

Generative models for images have seen tremendous progress in recent years. State-of-the-art text-to-image (T2I) diffusion models are now capable of synthesizing high-resolution photo-realistic images that adhere to complex text prompts (Saharia et al., 2022b; Ramesh et al., 2022; Rombach et al., 2022), and allow a wide range of image editing capabilities (Po et al., 2023) and other downstream uses. However, training large-scale text-to-video (T2V) foundation models remains an open challenge due to the added complexities that motion introduces. Not only are we sensitive to errors in modeling natural motion, but the added temporal data dimension introduces significant challenges in terms of memory and compute requirements, as well as the scale of the required training data to learn this more complex distribution. As a result, while T2V models are rapidly improving, existing models are still restricted in terms of video duration, overall visual quality, and the degree of realistic motion that they can generate.

A prevalent approach among existing T2V models is to adopt a cascaded design in which a base model generates distant keyframes, and subsequent temporal super-resolution (TSR) models generate the missing data between

the keyframes in non-overlapping segments. While memory efficient, the ability to generate globally coherent motion using temporal cascades is inherently restricted for the following reasons: (i) The base model generates an aggressively sub-sampled set of keyframes, in which fast motion becomes temporally aliased and thus ambiguous. (ii) TSR modules are constrained to fixed, small temporal context windows, and thus cannot consistently resolve aliasing ambiguities across the full duration of the video (illustrated in Fig. 2 in the case of synthesizing periodic motion, e.g., walking). (iii) Cascaded training regimens in general suffer from a domain gap, where the TSR model is trained on real downsampled video frames, but at inference time is used to interpolate generated frames, which accumulates errors.

Here, we take a different approach by introducing a new T2V diffusion framework that generates the full temporal duration of the video *at once*. We achieve this by using a Space-Time U-Net (STUNet) architecture that learns to downsample the signal in both space *and time*, and performs the majority of its computation in a compact space-time representation. This approach allows us to generate 80 frames at 16fps (or 5 seconds, which is longer than the average shot duration in most media (Cutting & Candan,

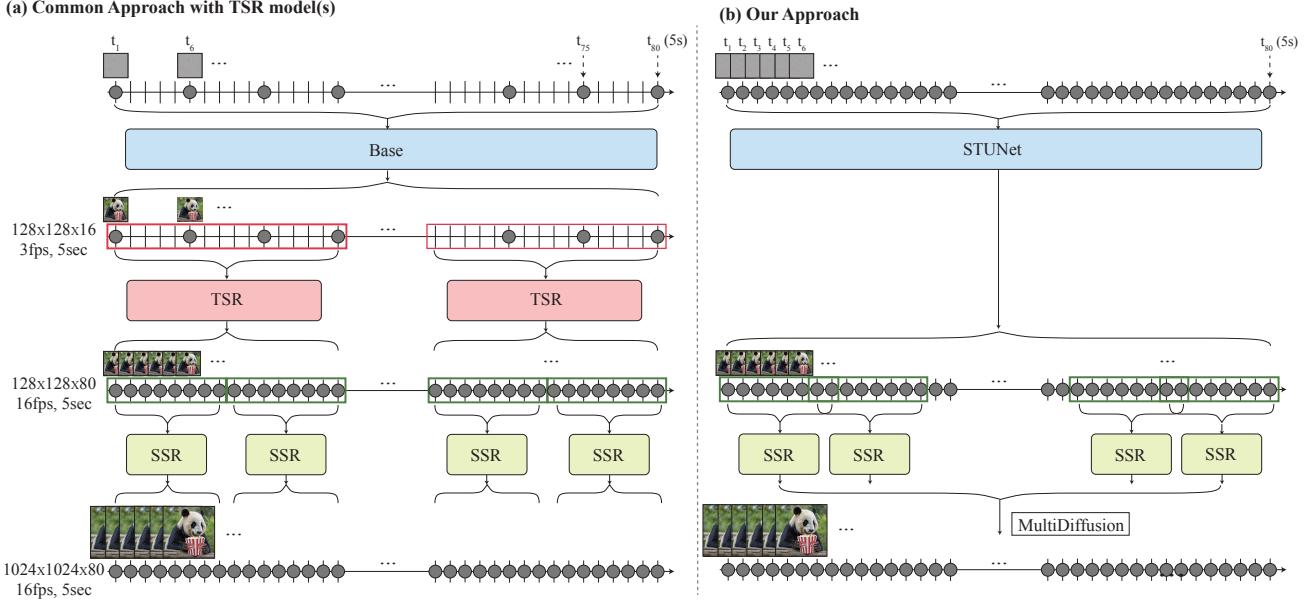


Figure 3: Lumiere pipeline. We illustrate our pipeline and the main difference from most common approach taken by previous works. (a) The common approach consists of a base model that generates distant keyframes, and a cascade of temporal super-resolution (TSR) models which subsequently fill in frames. A spatial super-resolution (SSR) model is applied in non-overlapping windows to obtain the high-resolution result. (b) In contrast, the base model in our framework process all frames *at once*, without a cascade of TSR models, allowing us to learn globally coherent motion. To obtain the high-resolution video, we apply a SSR model on *overlapping* windows and utilize MultiDiffusion (Bar-Tal et al., 2023) to combine the predictions into a coherent result. See Sec.3 for details.

2015)) with a single base model, leading to more globally coherent motion compared to prior work. Surprisingly, this design choice has been overlooked by previous T2V models, which follow the convention to include *only spatial* down-and up-sampling operations in the architecture, and maintain a *fixed temporal resolution* across the network (Ho et al., 2022b;a; Singer et al., 2022; Ge et al., 2023; Blattmann et al., 2023b; Wang et al., 2023a; Guo et al., 2023; Zhang et al., 2023a; Girdhar et al., 2023; Po et al., 2023).

To benefit from the powerful generative prior of T2I models, we follow the trend of building Lumiere on top of a pretrained (and fixed) T2I model (Hong et al., 2022; Singer et al., 2022; Saharia et al., 2022b). In our case, the T2I model works in pixel space and consists of a base model followed by a *spatial* super-resolution (SSR) cascade. Since the SSR network operates at high spatial resolution, applying it on the entire video duration is infeasible in terms of memory requirements. Common SSR solutions use a temporal windowing approach, which splits the video into non-overlapping segments and stitches together the results. However, this can lead to inconsistencies in appearance at the boundaries between windows (Girdhar et al., 2023). We propose to extend Multidiffusion (Bar-Tal et al., 2023), an approach proposed for achieving global continuity in panoramic image generation, to the temporal domain, where

we compute spatial super-resolution on temporal windows, and aggregate results into a globally coherent solution over the whole video clip.

We demonstrate state-of-the-art video generation results and show how to easily adapt Lumiere to a plethora of video content creation tasks, including video inpainting (Fig. 7), image-to-video generation (Fig. 5), or generating stylized videos that comply with a given style image (Fig. 6). Finally, we demonstrate that generating the full video at once allows us to easily invoke off-the-shelf editing methods to perform consistent editing (Fig. 9).

2. Related work

Text-to-Image Generation. Most of the common approaches for text-to-image (T2I) generation are based on diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020). Of these, DALL-E2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022b) achieve photorealistic text-to-image generation using cascaded diffusion models, whereas Stable Diffusion (Rombach et al., 2022) performs generation in a compressed low-dimensional latent space. A promising line of works design T2I diffusion models that generate high-resolution images end-to-end, without a spatial super-resolution cascaded system or fixed pre-trained

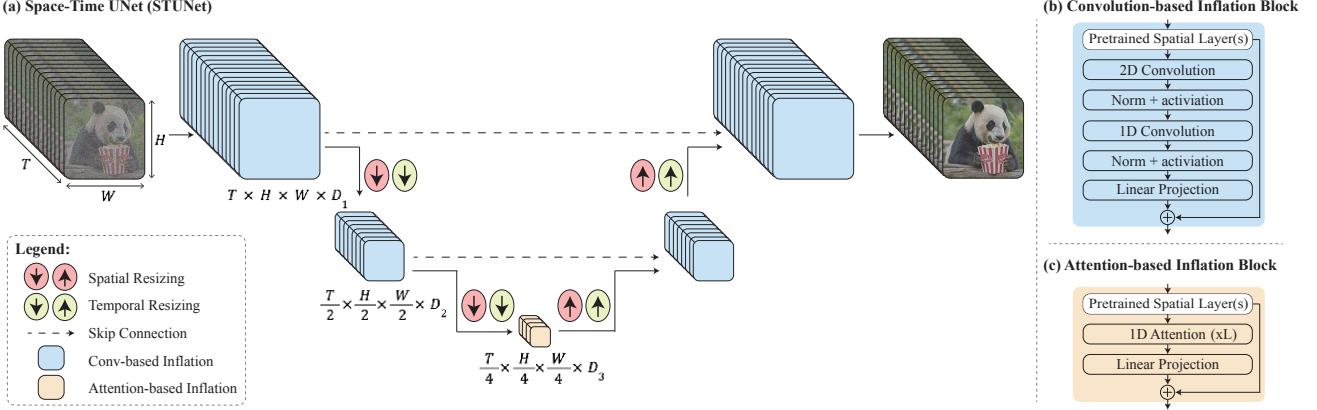


Figure 4: **STUNet architecture.** We “inflate” a pre-trained T2I U-Net architecture (Ho et al., 2022a) into a Space-Time UNet (STUNet) that down- and up-sample the video in both space and time. (a) An illustration of STUNet’s activation maps; color indicates features resulting from different temporal modules: (b) *Convolution-based* blocks which consist of pre-trained T2I layers followed by a factorized space-time convolution, and (c) *Attention-based* blocks at the coarsest U-Net level in which the pre-trained T2I layers are followed by temporal attention. Since the video representation is compressed at the coarsest level, we stack several temporal attention layers with limited computational overhead. See Sec. 3.1 for details.

latent space (Hoogeboom et al., 2023; Gu et al., 2023; Chen, 2023). Here, we design a T2V model that generates the full frame duration at once, avoiding the *temporal cascade* commonly involved in T2V models.

Text-to-Video Generation. Recently, there have been substantial efforts in training large-scale T2V models on large scale datasets with autoregressive Transformers (e.g., (Villegas et al., 2023; Wu et al., 2022; Hong et al., 2022; Kondratyuk et al., 2023)) or Diffusion Models (e.g., (Ho et al., 2022a;b)). A prominent approach for T2V generation is to “inflate” a pre-trained T2I model by inserting temporal layers to its architecture, and fine-tuning only those, or optionally the whole model, on video data (Singer et al., 2022; Blattmann et al., 2023b; Girdhar et al., 2023; Ge et al., 2023; Yuan et al., 2024). PYoCo (Ge et al., 2023) carefully design video noise prior and obtain better performance for fine-tuning a T2I model for video generation. VideoOLDM (Blattmann et al., 2023b) and AnimateDiff (Guo et al., 2023) inflate StableDiffusion (Rombach et al., 2022) and train only the newly-added temporal layers, showing they can be combined with the weights of personalized T2I models. Interestingly, the ubiquitous convention of existing inflation schemes is to maintain a *fixed temporal resolution* across the network, which limits their ability to process full-length clips. In this work, we design a new inflation scheme which includes learning to downsample the video in both space *and time*, and performing the majority of computation in the compressed space-time feature space of the network. We extend an Imagen T2I (Saharia et al., 2022b), however our architectural contributions could be used for latent diffusion as well, and are orthogonal to possible improvements

to the diffusion noise scheduler (Ge et al., 2023) or to the video data curation (Blattmann et al., 2023a).

3. Lumiere

We utilize Diffusion Probabilistic Models as our generative approach (Sohl-Dickstein et al., 2015; Croitoru et al., 2023a; Dhariwal & Nichol, 2021; Ho et al., 2020; Nichol & Dhariwal, 2021). These models are trained to approximate a data distribution (in our case, a distribution over videos) through a series of denoising steps. Starting from a Gaussian i.i.d. noise sample, the diffusion model gradually denoises it until reaching a clean sample drawn from the approximated target distribution. Diffusion models can learn a conditional distribution by incorporating additional guiding signals, such as text embedding, or spatial conditioning (e.g., depth map) (Dhariwal & Nichol, 2021; Saharia et al., 2022a; Croitoru et al., 2023b; Zhang et al., 2023b).

Our framework consists of a base model and a spatial super-resolution (SSR) model. As illustrated in Fig. 3b, our base model generates full clips at a coarse spatial resolution. The output of our base model is spatially upsampled using a temporally-aware SSR model, resulting with the high-resolution video. We next describe the key design choices in our architecture, and demonstrate the applicability of our framework for a variety of downstream applications.

3.1. Space-Time U-Net (STUnet)

To make our problem computationally tractable, we propose to use a space-time U-Net which downsamples the input signal both spatially *and* temporally, and performs the majority

Text-to-Video

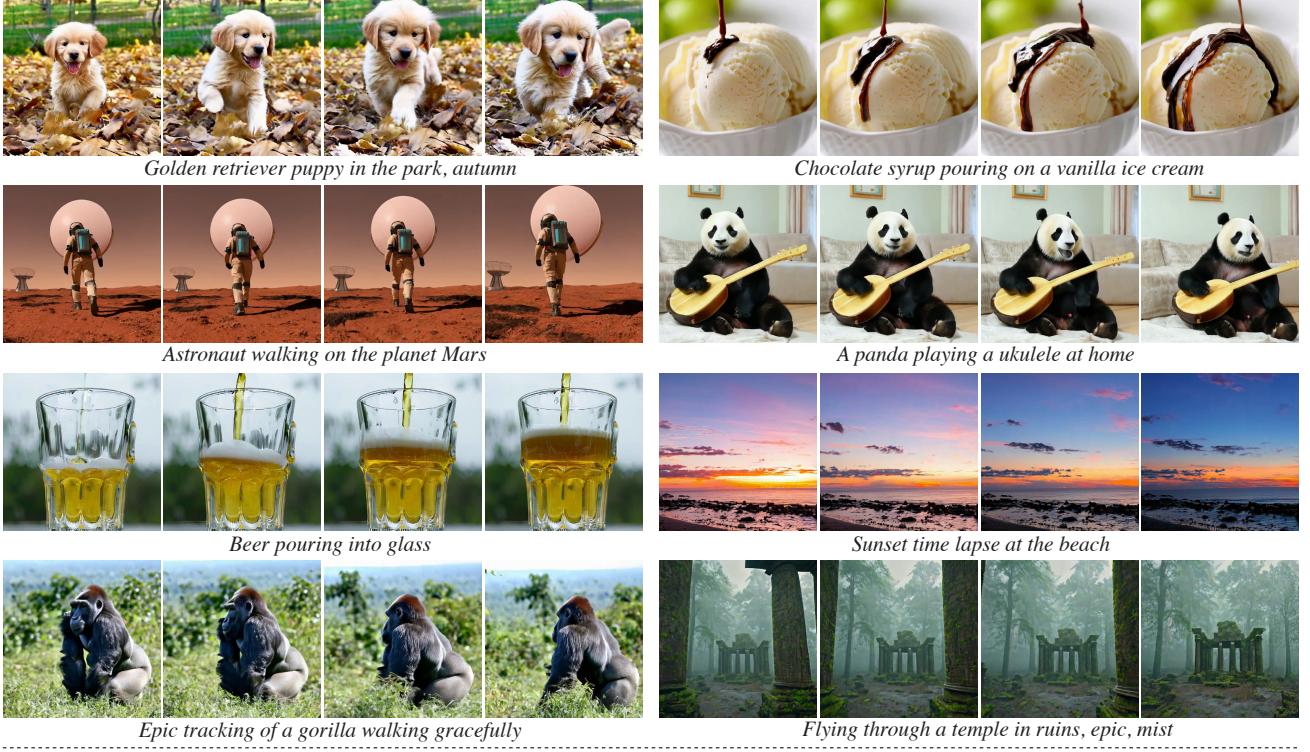


Image-to-Video

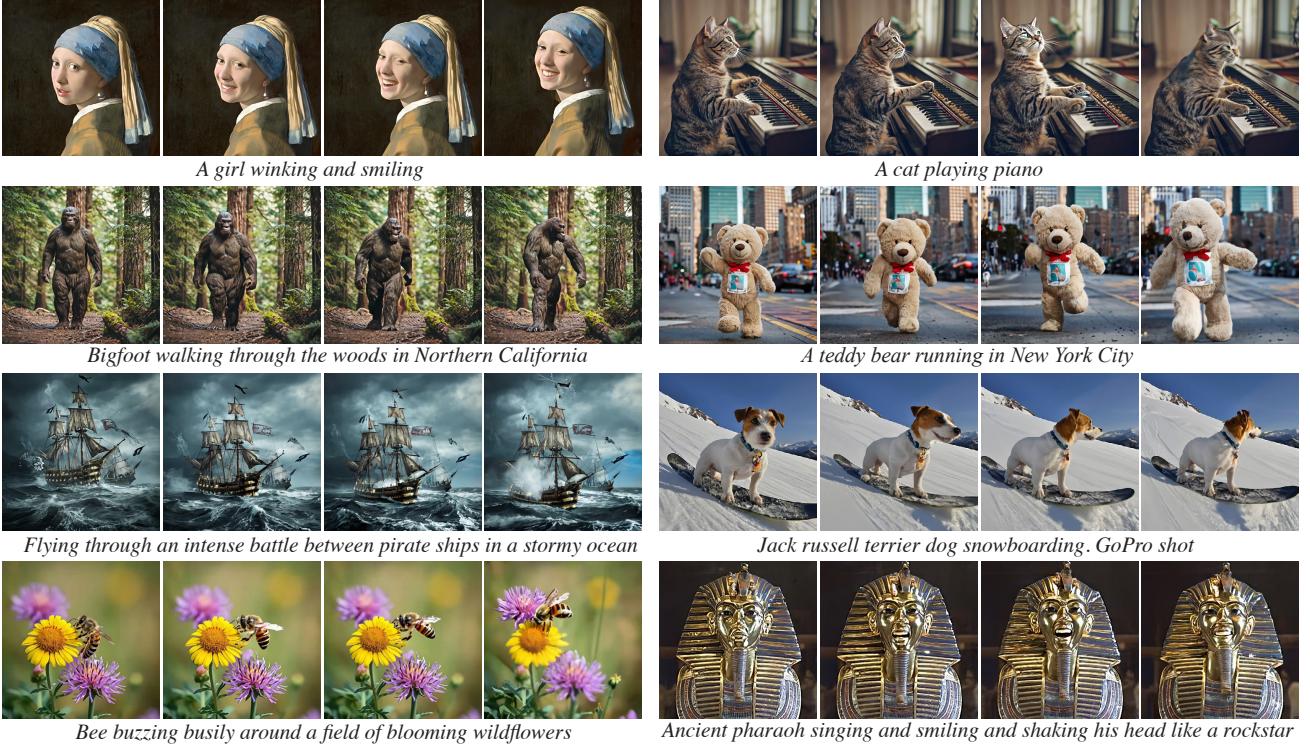


Figure 5: **Video generation results.** Sample results of text-to-video and image-to-video generation. The text prompt is indicated below each example. For image-to-video, the leftmost frame is provided to the model as a condition (see Sec. 4.2). We refer the reader to the SM for full-video results.

of its computation on this compact space-time representation. We draw inspiration from Çiçek et al. (2016), who generalize the U-Net architecture (Ronneberger et al., 2015) to include 3D pooling operations for efficient processing of volumetric biomedical data.

Our architecture is illustrated in Fig. 4. We interleave temporal blocks in the T2I architecture, and insert temporal down- and up-sampling modules following each pre-trained spatial resizing module (Fig. 4a). The temporal blocks include temporal convolutions (Fig. 4b) and temporal attention (Fig. 4c). Specifically, in all levels except for the coarsest, we insert factorized space-time convolutions (Fig. 4b) which allow increasing the non-linearities in the network compared to full-3D convolutions while reducing the computational costs, and increasing the expressiveness compared to 1D convolutions (Tran et al., 2018). As the computational requirements of temporal attention scale quadratically with the number of frames, we incorporate temporal attention only at the coarsest resolution, which contains a space-time compressed representation of the video. Operating on the low dimensional feature map allows us to stack several temporal attention blocks with limited computational overhead.

Similarly to (Blattmann et al., 2023b; Guo et al., 2023), we train the newly added parameters, and keep the weights of the pre-trained T2I fixed. Notably, the common inflation approach ensures that at initialization, the T2V model is equivalent to the pre-trained T2I model, i.e., generates videos as a collection of independent image samples. However, in our case, it is impossible to satisfy this property due to the temporal down- and up-sampling modules. We empirically found that initializing these modules such that they perform nearest-neighbor down- and up-sampling operations results with a good starting point (in terms of the loss function).

3.2. Multidiffusion for Spatial-Super Resolution

Due to memory constraints, the inflated SSR network can operate only on short segments of the video. To avoid temporal boundary artifacts, we achieve smooth transitions between the temporal segments by employing Multidiffusion (Bar-Tal et al., 2023) along the temporal axis. At each generation step, we split the noisy input video $J \in \mathbb{R}^{H \times W \times T \times 3}$ into a set of overlapping segments $\{J_i\}_{i=1}^N$, where $J_i \in \mathbb{R}^{H \times W \times T' \times 3}$ is the i^{th} segment, which has temporal duration $T' < T$. To reconcile the per-segment SSR predictions, $\{\Phi(J_i)\}_{i=1}^N$, we define the result of the denoising step to be the solution of the optimization problem

$$\arg \min_{J'} \sum_{i=1}^n \|J' - \Phi(J_i)\|^2.$$

The solution to this problem is given by linearly combining the predictions over overlapping windows.

4. Applications

The lack of a TSR cascade makes it easier to extend Lumiere to downstream applications. In particular, our model provides an intuitive interface for downstream applications that require an off-the-shelf T2V model (e.g., Meng et al. (2022); Poole et al. (2023); Gal et al. (2023)). We demonstrate this property by performing video-to-video editing using SDEdit (Meng et al., 2022) (see Fig. 9). We next discuss a number of such applications, including style conditioned generation, image-to-video, inpainting and outpainting, and cinemagraphs. We present example frames in Figs. 6–9 and refer the reader to the Supplementary Material (SM) on our webpage for full video results.

4.1. Stylized Generation

Recall that we only train the newly-added temporal layers and keep the pre-trained T2I weights fixed. Previous work showed that substituting the T2I weights with a model customized for a specific style allows to generate videos with the desired style (Guo et al., 2023). We observe that this simple “plug-and-play” approach often results in distorted or static videos (see SM), and hypothesize that this is caused by the significant deviation in the distribution of the input to the temporal layers from the fine-tuned spatial layers.

Inspired by the success of GAN-based interpolation approaches (Pinkney & Adler, 2020), we opt to strike a balance between style and motion by linearly interpolating between the fine-tuned T2I weights, W_{style} , and the original T2I weights, W_{orig} . Specifically, we construct the interpolated weights as $W_{\text{interpolate}} = \alpha \cdot W_{\text{style}} + (1 - \alpha) \cdot W_{\text{orig}}$. The interpolation coefficient $\alpha \in [0.5, 1]$ is chosen manually in our experiments to generate videos that adhere to the style and depict plausible motion.

Figure 6 presents sample results for various styles from (Sohn et al., 2023). While more realistic styles such as “watercolor painting” result in realistic motion, other, less realistic spatial priors derived from vector art styles, result in corresponding unique non-realistic motion. For example, the “line drawing” style results in animations that resemble pencil strokes “drawing” the described scene, while the “cartoon” style result in content that gradually “pops out” and constructs the scene (see SM for full videos).

4.2. Conditional Generation

Similarly to Blattmann et al. (2023b); Wang et al. (2023b), we extend our model to video generation conditioned on additional input signals (e.g., image or mask). We achieve this by modifying the model to take as input two signals in addition to the noisy video $J \in \mathbb{R}^{T \times H \times W \times 3}$ and the driving text prompt. Specifically, we add the masked conditioning video $C \in \mathbb{R}^{T \times H \times W \times 3}$ and its corresponding binary mask

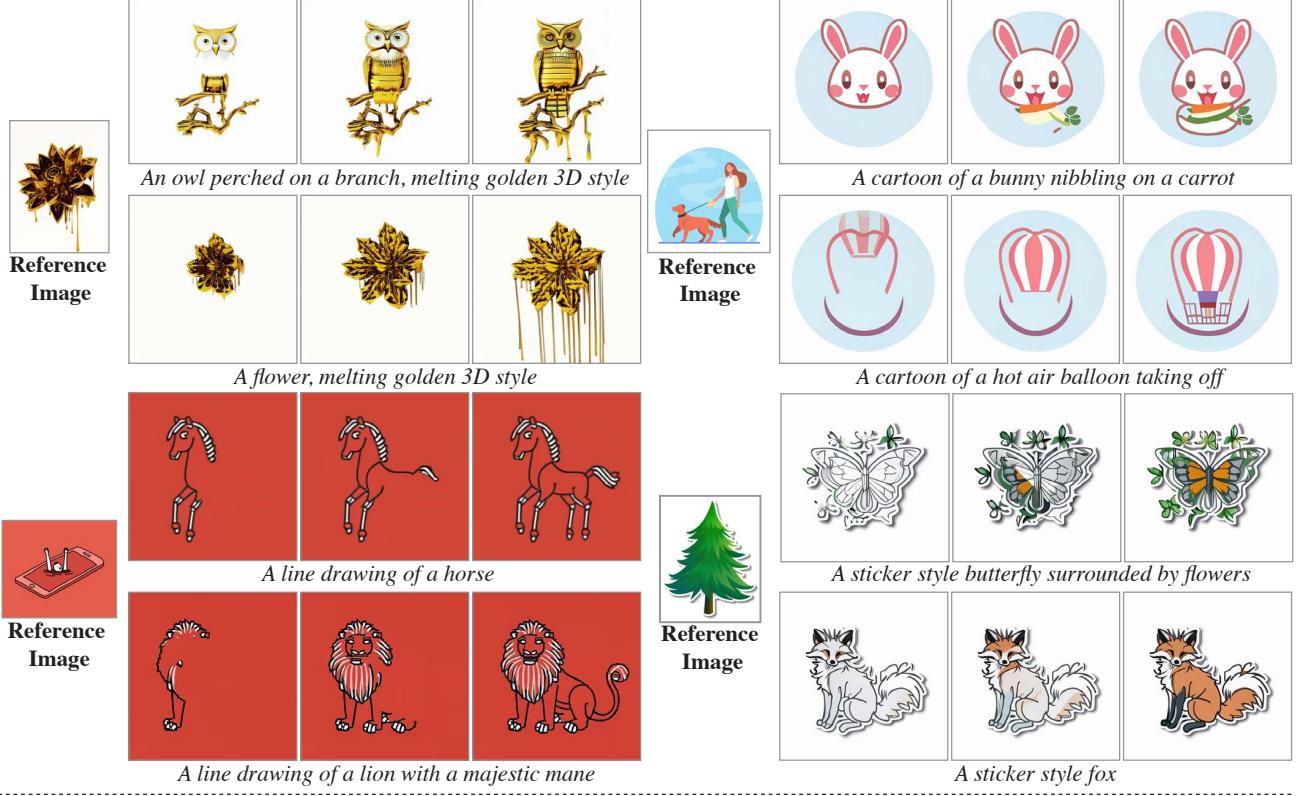
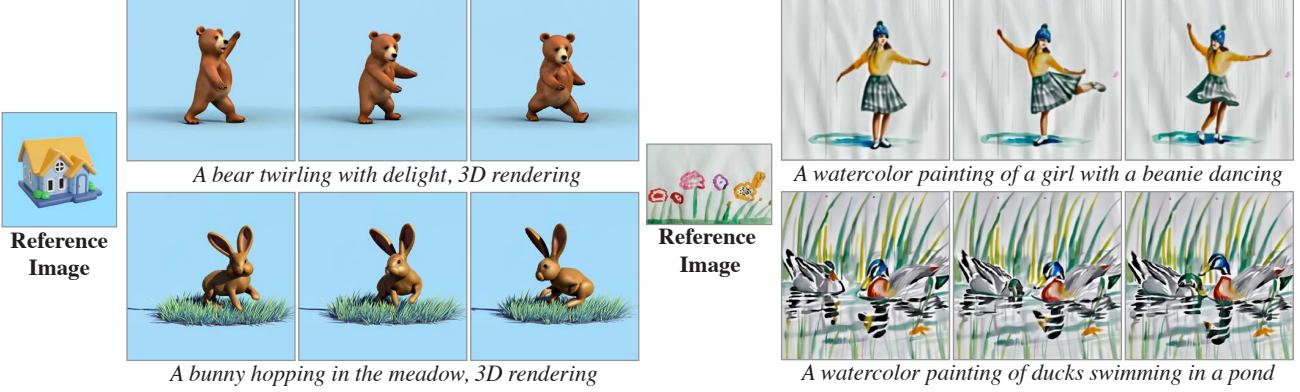
(A)

(B)


Figure 6: Stylized Generation. Given a driving style image and its corresponding set of fine-tuned text-to-image weights, we perform linear interpolation between the fine-tuned and pre-trained weights of the model’s spatial layers. We present results for (A) vector art styles, and (B) realistic styles. The results demonstrate Lumiere’s ability to creatively match a *different* motion prior to each of the spatial styles (frames shown from left to right). See Sec. 4.1 for details.

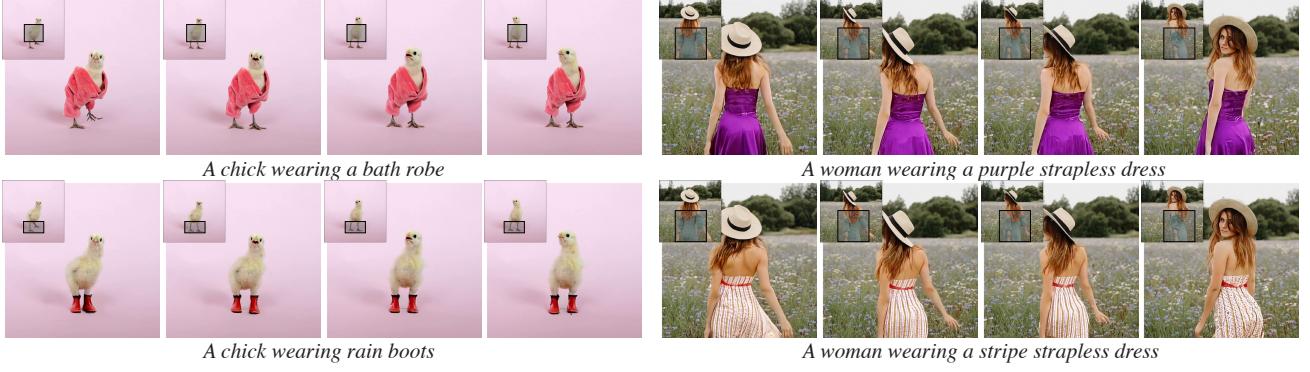


Figure 7: **Inpainting.** Examples of video inpainting with Lumiere. For each input video (top left corner of each frame), we animate the masked area of the video using our model.

$M \in \mathbb{R}^{T \times H \times W \times 1}$, such that the overall input to the model is the concatenated tensor $\langle J, C, M \rangle \in \mathbb{R}^{T \times H \times W \times 7}$. We expand the channel dimension of the first convolution layer from 3 to 7 in order to accommodate the modified input shape and fine-tune our base T2V model to denoise J based on C, M . During this fine-tuning process, we take J to be the noisy version of the training video, and C to be a masked version of the clean video. This encourages the model to learn to copy the unmasked information in C to the output video while only animating the masked content, as desired.

Image-to-Video. In this case, the first frame of the video is given as input. The conditioning signal C contains this first frame followed by blank frames for the rest of the video. The corresponding mask M contains ones (i.e., unmasked content) for the first frame and zeros (i.e., masked content) for the rest of the video. Figures 1 and 5 show sample results of image-conditioned generation (see SM for more results). Our model generates videos that start with the desired first frame, and exhibit intricate coherent motion across the entire video duration.

Inpainting. Here, the conditioning signals are a user-provided video C and a mask M that describes the region to complete in the video. Note that the inpainting application can be used for object replacement/insertion (Fig. 1) as well as for localized editing (Fig. 7). The effect is a seamless and natural completion of the masked region, with contents guided by the text prompt. We refer the reader to the SM for more examples of both inpainting and outpainting.

Cinemagraphs. We additionally consider the application of animating the content of *an image* only within a specific user-provided region. The conditioning signal C is the input image duplicated across the entire video, while the mask M contains ones for the entire first frame (i.e., the first frame is unmasked), and for the other frames, the mask contains ones only outside the user-provided region (i.e.,

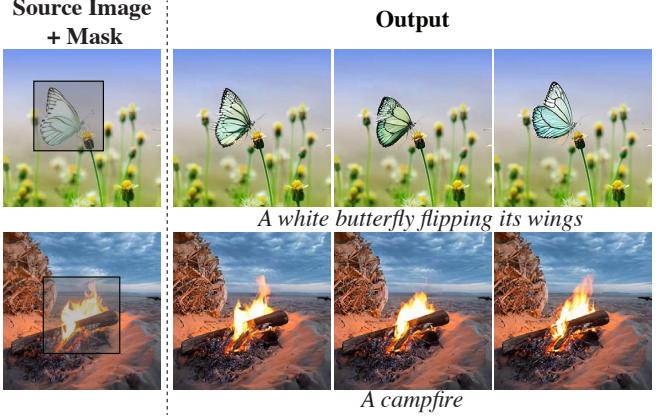


Figure 8: **Cinemagraphs.** Given only an input *image* and a mask (left), our method generates a video in which the marked area is animated and the rest remains static (right).

the other frames are masked inside the region we wish to animate). We provide sample results in Fig. 8 and in the SM. Since the first frame remains unmasked, the animated content is encouraged to maintain the appearance from the conditioning image.

5. Evaluation and Comparisons

We train our T2V model on a dataset containing 30M videos along with their text caption. The videos are 80 frames long at 16 fps (5 seconds). The base model is trained at 128×128 and the SSR outputs 1024×1024 frames. We evaluate our model on a collection of 113 text prompts describing diverse objects and scenes. The prompt list consists of 18 prompts assembled by us and 95 prompts used by prior works (Singer et al., 2022; Ho et al., 2022a; Blattmann et al., 2023b) (see App. B). Additionally, we employ a zero-shot evaluation protocol on the UCF101 dataset (Soomro et al., 2012), as detailed in Sec. 5.2.

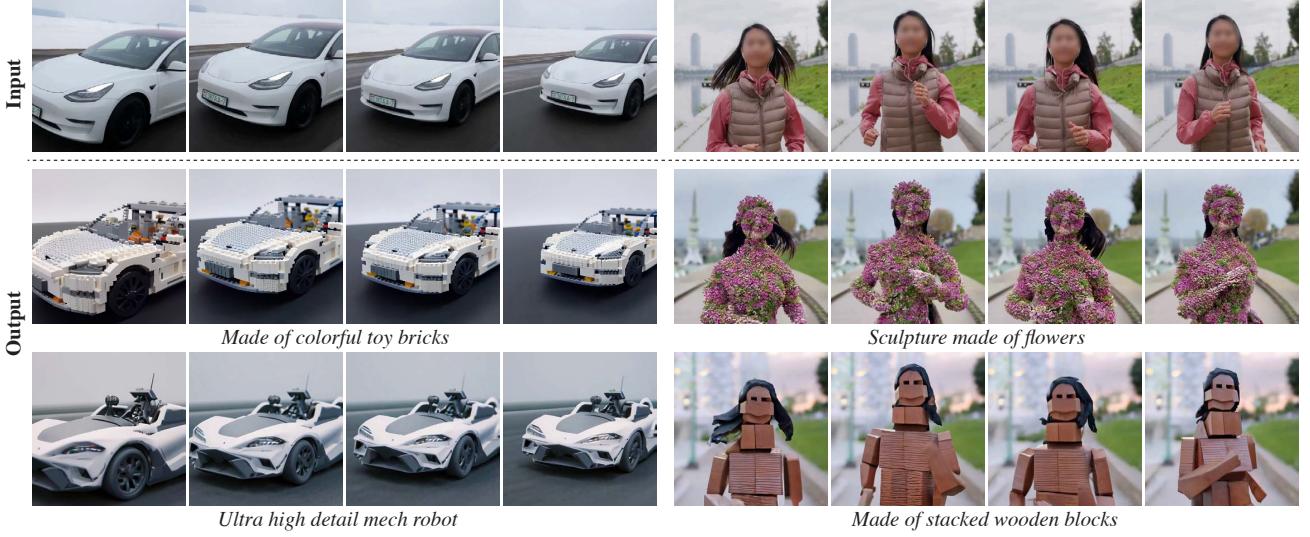


Figure 9: **Video-to-video via SDEdit.** Our base model generates full-frame-rate videos, without a TSR cascade, and thus facilitates an intuitive interface for downstream applications. We demonstrate this property by applying SDEdit (Meng et al., 2022) using our model, achieving consistent video stylization. We show several frames of a given input video in the first row, and the corresponding edited frames below.

Method	FVD ↓	IS ↑
MagicVideo (Zhou et al., 2022)	655.00	-
Emu Video (Girdhar et al., 2023)	606.20	42.70
Video LDM (Blattmann et al., 2023b)	550.61	33.45
Show-1 (Zhang et al., 2023a)	394.46	35.42
Make-A-Video (Singer et al., 2022)	367.23	33.00
PYCo (Ge et al., 2023)	355.19	47.76
SVD (Blattmann et al., 2023a)	242.02	-
Lumiere (Ours)	332.49	37.54

Table 1: Zero-shot text-to-video generation comparison on UCF101 (Soomro et al., 2012). Our method achieves competitive FVD (Unterthiner et al., 2018) and IS (Salimans et al., 2016) scores. See Sec. 5.2.

We illustrate text-to-video generation in Figs. 1 and 5. Our method generates high-quality videos depicting both intricate object motion (e.g., walking astronaut in Fig. 5) and coherent camera motion (e.g., car example in Fig. 1). We refer the reader to the SM for full-video results.

Baselines. We compare our method to prominent T2V diffusion models: (i) ImagenVideo (Ho et al., 2022a), that operates in pixel-space and consists of a cascade of 7 models (a base model, 3 TSR models, and 3 SSR models); (ii) AnimateDiff (Guo et al., 2023), (iii) StableVideoDiffusion (SVD) (Blattmann et al., 2023a), and (iv) ZeroScope (Wang et al., 2023a) that inflate Stable Diffusion (Rombach et al., 2022) and train on video data; note that AnimateDiff and ZeroScope output only 16, and 36 frames respectively. SVD

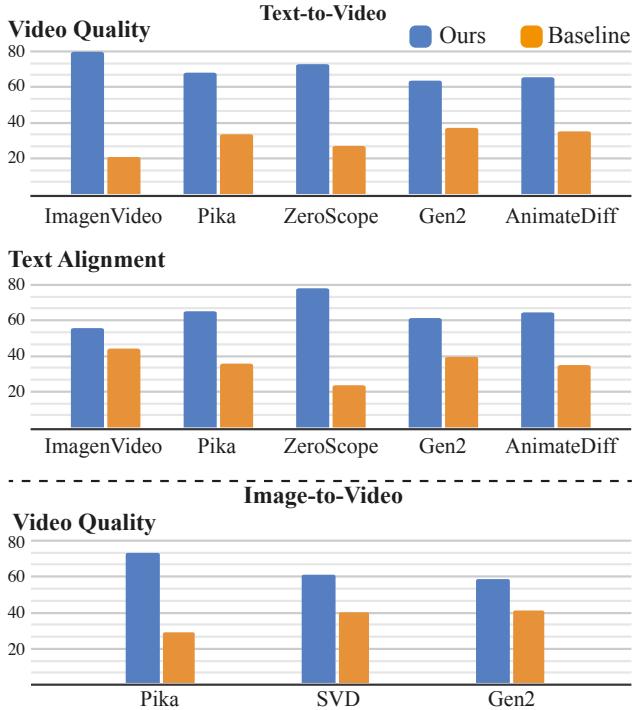


Figure 10: **User study results.** Our method was preferred by users in both text-to-video, and image-to-video generation. See Sec. 5.2.

released only their image-to-video model, which outputs 25 frames and is not conditioned on text. Additionally, we compare to (v) Pika ([Pika labs, 2023](#)) and (vi) Gen2 ([RunwayML, 2023](#)) commercial T2V models that have available API. Furthermore, we qualitatively compare to additional T2V models that are closed-source in Sec. 5.2.

5.1. Qualitative Evaluation

We provide qualitative comparison between our model and the baselines in Fig. 11. We observed that Gen-2 ([RunwayML, 2023](#)) and Pika ([Pika labs, 2023](#)) demonstrate high per-frame visual quality; however, their outputs are characterized by a very limited amount of motion, often resulting in near-static videos. ImagenVideo ([Ho et al., 2022a](#)) produces a reasonable amount of motion, but at a lower overall visual quality. AnimateDiff ([Guo et al., 2023](#)) and ZeroScope ([Wang et al., 2023a](#)) exhibit noticeable motion but are also prone to visual artifacts. Moreover, they generate videos of shorter durations, specifically 2 seconds and 3.6 seconds, respectively. In contrast, our method produces 5-second videos that have higher motion magnitude while maintaining temporal consistency and overall quality.

5.2. Quantitative Evaluation

Zero-shot evaluation on UCF101. Following the evaluation protocols of [Blattmann et al. \(2023a\)](#) and [Ge et al. \(2023\)](#), we quantitatively evaluate our method for zero-shot text-to-video generation on UCF101 ([Soomro et al., 2012](#)). Table 1 reports the Fréchet Video Distance (FVD) ([Unterthiner et al., 2018](#)) and Inception Score (IS) ([Salimans et al., 2016](#)) of our method and previous work. We achieve competitive FVD and IS scores. However, as discussed in previous work (e.g., [Girdhar et al. \(2023\)](#); [Ho et al. \(2022a\)](#); [Chong & Forsyth \(2020\)](#)), these metrics do not faithfully reflect human perception, and may be significantly influenced by low-level details ([Parmar et al., 2022](#)) and by the distribution shift between the reference UCF101 data and the T2V training data ([Girdhar et al., 2023](#)). Furthermore, the protocol uses only 16 frames from generated videos and thus is not able to capture long-term motion.

User Study. We adopt the Two-alternative Forced Choice (2AFC) protocol, as used in previous works ([Kolkin et al., 2019](#); [Zhang et al., 2018](#); [Blattmann et al., 2023a](#); [Rombach et al., 2022](#)). In this protocol, participants were presented with a randomly selected pair of videos: one generated by our model and the other by one of the baseline methods. Participants were then asked to choose the video they deemed better in terms of visual quality and motion. Additionally, they were asked to select the video that more accurately matched the target text prompt. We collected ~ 400 user judgments for each baseline and question, utilizing the Amazon Mechanical Turk (AMT) platform. As illustrated in

Fig. 10, our method was preferred over all baselines by the users and demonstrated better alignment with the text prompts. Note that ZeroScope and AnimateDiff generate videos only at 3.6 and 2 second respectively, we thus trim our videos to match their duration when comparing to them.

We further conduct a user study for comparing our image-to-video model (see Sec. 4.2) against Pika ([Pika labs, 2023](#)), StableVideoDiffusion (SVD) ([Blattmann et al., 2023a](#)), and Gen2([RunwayML, 2023](#)). Note that SVD image-to-video model is not conditioned on text, we thus focus our survey on the video quality. As seen in Fig. 10, our method was preferred by users compared to the baselines. For a detailed description of the full evaluation protocol, please refer to Appendix B.

6. Conclusion

We presented a new text-to-video generation framework, utilizing a pre-trained text-to-image diffusion model. We identified an inherent limitation in learning globally-coherent motion in the prevalent approach of first generating *distant* keyframes and subsequently interpolating them using a cascade of temporal super-resolution models. To tackle this challenge, we introduced a space-time U-Net architecture design that directly generates full-frame-rate video clips, by incorporating both spatial, *and* temporal down- and up-sampling modules. We demonstrated state-of-the-art generation results, and showed the applicability of our approach for a wide range of applications, including image-to-video, video inpainting, and stylized generation.

As for limitations, our method is not designed to generate videos that consist of multiple shots, or that involve transitions between scenes. Generating such content remains an open challenge for future research. Furthermore, we established our model on top of a T2I model that operates in the pixel space, and thus involves a spatial super resolution module to produce high resolution images. Nevertheless, our design principles are applicable to latent video diffusion models ([Rombach et al., 2022](#)), and can trigger further research in the design of text-to-video models.

7. Societal Impact

Our primary goal in this work is to enable novice users to generate visual content in an creative and flexible way. However, there is a risk of misuse for creating fake or harmful content with our technology, and we believe that it is crucial to develop and apply tools for detecting biases and malicious use cases in order to ensure a safe and fair use.

Acknowledgments We would like to thank Miki Rubinstein, Amit Raj, Ronny Votel, Orly Liba, Bryan Seybold, David Ross, Guanghui Liu, Dan Goldman, Hartwig Adam,

Xuhui Jia, Xiuye Gu, Rachel Hornung, Oran Lang and Jess Gallegos for their collaboration, helpful discussions, feedback and support. We thank owners of images and videos used in our experiments for sharing their valuable assets (attributions can be found in our webpage).

References

- Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. Multi-Diffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023.
- 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*, 2023a.
- Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.
- Carreira, J. and Zisserman, A. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, pp. 6299–6308, 2017.
- Chen, T. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.
- Chong, M. J. and Forsyth, D. Effectively unbiased FID and Inception Score and where to find them. In *CVPR*, pp. 6070–6079, 2020.
- Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *MICCAI*, pp. 424–432. Springer, 2016.
- Croitoru, F.-A., Hondu, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE T. Pattern Anal. Mach. Intell.*, 2023a.
- Croitoru, F.-A., Hondu, V., Ionescu, R. T., and Shah, M. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023b.
- Cutting, J. E. and Candan, A. Shot durations, shot classes, and the increased pace of popular movies, 2015.
- Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *NeurIPS*, 2021.
- Gal, R., Vinker, Y., Alaluf, Y., Bermano, A. H., Cohen-Or, D., Shamir, A., and Chechik, G. Breathing life into sketches using text-to-video priors. *arXiv preprint arXiv:2311.13608*, 2023.
- Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.-B., Liu, M.-Y., and Balaji, Y. Preserve your own correlation: A noise prior for video diffusion models. In *ICCV*, pp. 22930–22941, 2023.
- Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S. S., Shah, A., Yin, X., Parikh, D., and Misra, I. Emu Video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.
- Gu, J., Zhai, S., Zhang, Y., Susskind, J., and Jaitly, N. Matryoshka diffusion models. *arXiv preprint arXiv:2310.15111*, 2023.
- Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020.
- Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.
- Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models, 2022b.
- Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- Hoogeboom, E., Heek, J., and Salimans, T. Simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 2023.
- Kolkin, N., Salavon, J., and Shakhnarovich, G. Style transfer by relaxed optimal transport and self-similarity. In *CVPR*, pp. 10051–10060, 2019.
- Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al. VideoPoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.
- Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.
- Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *ICML*, pp. 8162–8171, 2021.

- Parmar, G., Zhang, R., and Zhu, J.-Y. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.
- Pika labs. <https://www.pika.art/>, 2023.
- Pinkney, J. N. and Adler, D. Resolution dependent GAN interpolation for controllable image synthesis between domains. In *Machine Learning for Creativity and Design NeurIPS 2020 Workshop*, 2020.
- Po, R., Yifan, W., Golyanik, V., Aberman, K., Barron, J. T., Bermano, A. H., Chan, E. R., Dekel, T., Holynski, A., Kanazawa, A., et al. State of the art on diffusion models for visual computing. *arXiv preprint arXiv:2310.07204*, 2023.
- Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. DreamFusion: Text-to-3D using 2D diffusion. In *ICLR*, 2023.
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.
- Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pp. 234–241. Springer, 2015.
- RunwayML. Gen-2. <https://research.runwayml.com/gen2>, 2023.
- Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35:36479–36494, 2022b.
- Saito, M., Saito, S., Koyama, M., and Kobayashi, S. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal GAN. *Int. J. Comput. Vision*, 128(10-11):2586–2606, 2020.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. *NIPS*, 29, 2016.
- Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-Video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265, 2015.
- Sohn, K., Ruiz, N., Lee, K., Chin, D. C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., et al. StyleDrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Soomro, K., Zamir, A. R., and Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pp. 6450–6459, 2018.
- Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.
- Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023.
- Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.
- Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., and Zhou, J. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023b.
- Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. Nüwa: Visual synthesis pre-training for neural visual world creation. In *ECCV*, pp. 720–736. Springer, 2022.
- Yuan, X., Baek, J., Xu, K., Tov, O., and Fei, H. Inflation with diffusion: Efficient temporal adaptation for text-to-video super-resolution, 2024.
- Zhang, D. J., Wu, J. Z., Liu, J.-W., Zhao, R., Ran, L., Gu, Y., Gao, D., and Shou, M. Z. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023a.
- Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., and Feng, J. MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

Appendix

A. Qualitative Comparison

We provide a qualitative comparison of our method and the baselines (see Sec. 5.1).

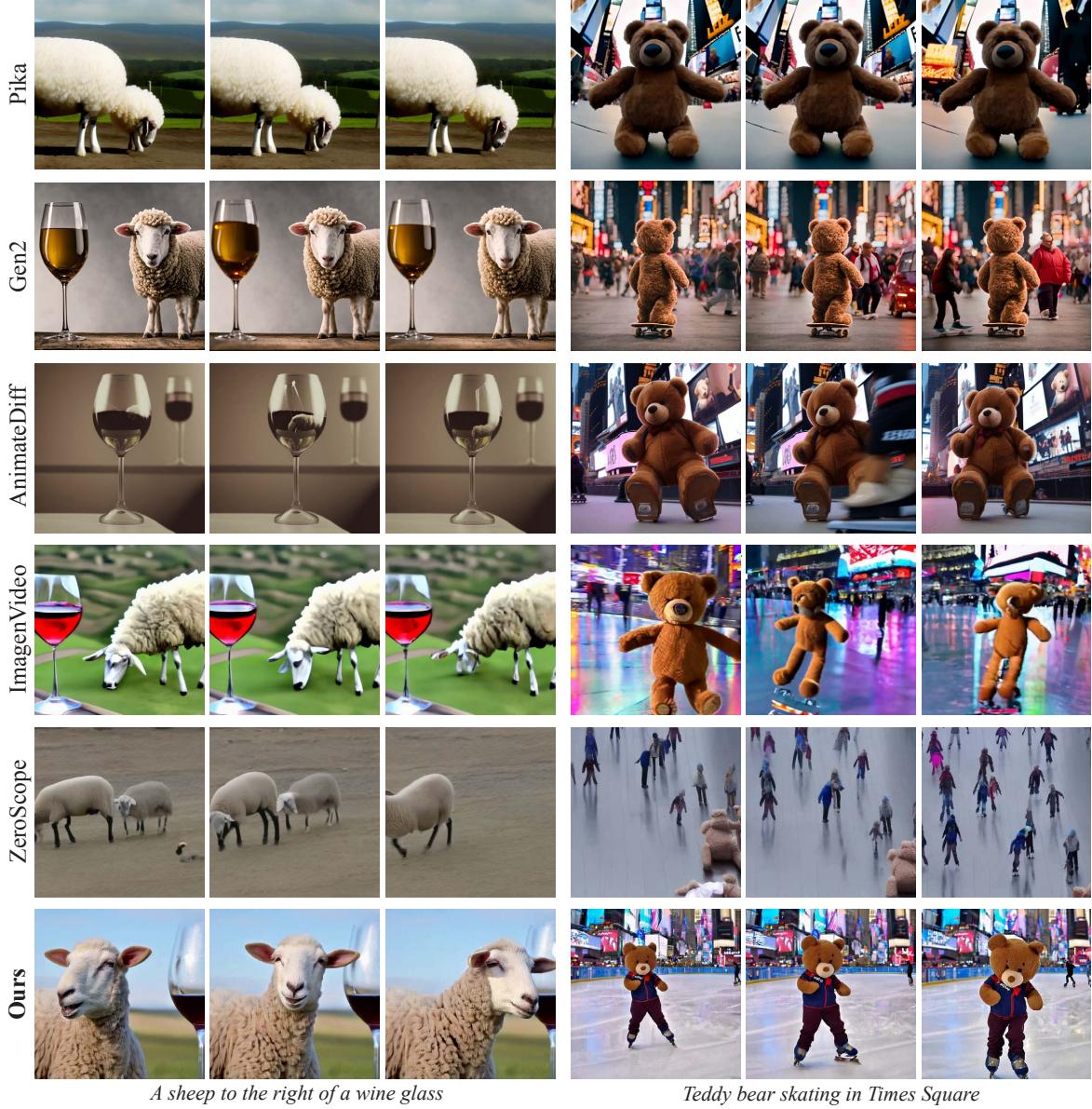


Figure 11: **Qualitative Comparison.** We qualitatively compare our model to prominent T2V diffusion models, Gen-2 (RunwayML, 2023), Pika (Pika labs, 2023), ImagenVideo (Ho et al., 2022a), AnimateDiff (Guo et al., 2023) and ZeroScope (Wang et al., 2023a).

B. Evaluation Protocol

B.1. User study

We used a set of 113 text prompts containing a variety of objects and actions for the text-to-video and image-to-video surveys. 95 prompts were taken from various recent text-to-video generation methods (Singer et al., 2022; Ho et al., 2022a; Blattmann et al., 2023b). The rest are prompts we created which describing complex scenes and actions, see Appendix B.3

for the full list of prompts. For each baseline method, we generated results for all prompts through their official APIs. To ensure a fair comparison, when generating results using our method we've fixed a random seed and generated all results using the same random seed without any curation. In addition, following (Girdhar et al., 2023) we align the spatial resolution and temporal duration of our videos to each one of the baselines. Spatially, we apply a central crop followed by resizing to 512×512 , and temporally, by trimming the start and the end such that the duration of both videos matches the shortest one. Each participant is shown 10 side-by-side comparisons between our method and a baseline (randomly ordered). We then asked the participant "Which video matches the following text?" and "Which video is better? Has more motion and better quality". We enable the user to answer only after watching the full video, waiting at least 3 seconds, and pressing a button reading "Ready to answer" (Fig. 12). To ensure credible responses, we include several vigilance tests during the study - a real video vs a complete random noise, and a static image vs a real video.

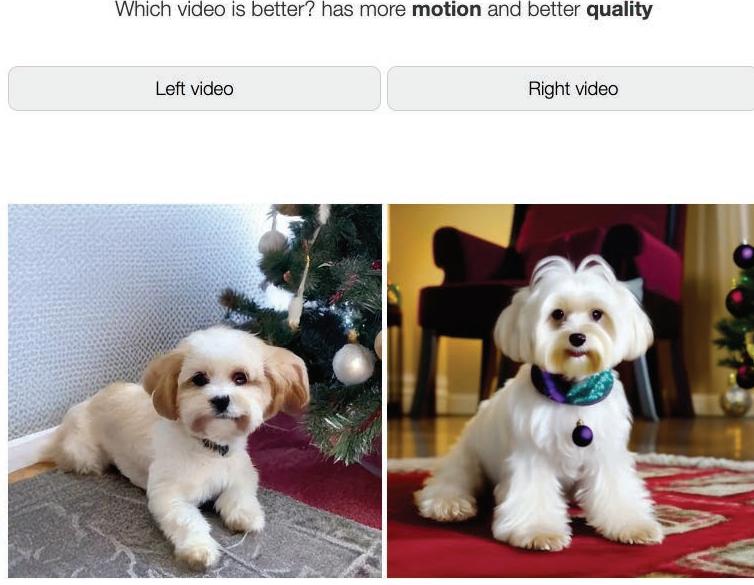


Figure 12: **User study display.** The user is shown two videos, ours and a baseline, randomly ordered.

B.2. Zero-shot evaluation on UCF101

To compute Fréchet Video Distance (FVD), we generated 10,235 videos by following the class distribution of the UCF101 dataset. We use the same prompt set from Ge et al. (2023) to generate the videos. After resizing all videos to 244×244 resolution, we extract I3D embedding (Carreira & Zisserman, 2017) of the first 16 frames of our videos. Then we compute the FVD score between the I3D embedding of ours and that of UCF101 videos. To compute Inception Score (IS), the same generated videos are used to extract C3D embedding (Saito et al., 2020).

B.3. List of prompts used for the user study evaluation

1. "A bear is giving a presentation in classroom."
2. "A bear wearing sunglasses and hosting a talk show."
3. "A beautiful sunrise on mars, Curiosity rover. High definition, timelapse, dramatic colors."
4. "A beautiful sunrise on mars. High definition, timelapse, dramatic colors."
5. "A bicycle on top of a boat."
6. "A big moon rises on top of Toronto city."
7. "A bigfoot walking in the snowstorm."
8. "A burning volcano."
9. "A car moving slowly on an empty street, rainy evening, van Gogh painting."

10. "A cat wearing sunglasses and working as a lifeguard at a pool."
11. "A confused grizzly bear in calculus class."
12. "A cute happy Corgi playing in park, sunset, 4k."
13. "A dog driving a car on a suburban street wearing funny sunglasses"
14. "A dog swimming."
15. "A dog wearing a Superhero outfit with red cape flying through the sky."
16. "A dog wearing virtual reality goggles in sunset, 4k, high resolution."
17. "A fantasy landscape, trending on artstation, 4k, high resolution."
18. "A fat rabbit wearing a purple robe walking through a fantasy landscape."
19. "A fire dragon breathing, trending on artstation, slow motion."
20. "A fox dressed in suit dancing in park."
21. "A giraffe underneath a microwave."
22. "A glass bead falling into water with a huge splash. Sunset in the background."
23. "A golden retriever has a picnic on a beautiful tropical beach at sunset, high resolution."
24. "A goldendoodle playing in a park by a lake."
25. "A hand lifts a cup."
26. "A happy elephant wearing a birthday hat walking under the sea."
27. "A horse galloping through van Gogh's Starry Night."
28. "A hot air balloon ascending over a lush valley, 4k, high definition."
29. "A koala bear playing piano in the forest."
30. "A lion standing on a surfboard in the ocean in sunset, 4k, high resolution."
31. "A monkey is playing bass guitar, stage background, 4k, high resolution."
32. "A panda standing on a surfboard in the ocean in sunset, 4k, high resolution."
33. "A panda taking a selfie."
34. "A polar bear is playing bass guitar in snow, 4k, high resolution."
35. "A raccoon dressed in suit playing the trumpet, stage background, 4k, high resolution."
36. "Red sports car coming around a bend in a mountain road."
37. "A shark swimming in clear Caribbean ocean."
38. "A sheep to the right of a wine glass."
39. "A shiny golden waterfall flowing through glacier at night."
40. "A shooting star flying through the night sky over mountains."
41. "A sloth playing video games and beating all the high scores."
42. "A small hand-crafted wooden boat taking off to space."
43. "A squirrel eating a burger."
44. "A storm trooper riding a bike in sunset street, 4k, high resolution."
45. "A storm trooper vacuuming the beach."
46. "A stunning aerial drone footage time lapse of El Capitan in Yosemite National Park at sunset."
47. "A swarm of bees flying around their hive."
48. "A teddy bear is playing the electric guitar, high definition, 4k."
49. "A teddy bear running in New York City."
50. "A teddy bear skating in Times Square."
51. "A video of the Earth rotating in space."
52. "A video that showcases the beauty of nature, from mountains and waterfalls to forests and oceans."
53. "Aerial view of a hiker man standing on a mountain peak."

54. "Aerial view of a lake in Switzerland"
55. "Aerial view of a snow-covered mountain."
56. "Aerial view of a white sandy beach on the shores of a beautiful sea, 4k, high resolution."
57. "An animated painting of fluffy white clouds moving in sky."
58. "An astronaut cooking with a pan and fire in the kitchen, high definition, 4k."
59. "An astronaut riding a horse, high definition, 4k."
60. "An astronaut riding a horse."
61. "An astronaut riding a horse in sunset, 4k, high resolution."
62. "Artistic silhouette of a wolf against a twilight sky, embodying the spirit of the wilderness.."
63. "Back view on young woman dressed in a yellow jacket walking in the forest."
64. "Maltipoo dog on the carpet next to a Christmas tree in a beautiful living room."
65. "Beer pouring into glass, low angle video shot."
66. "Bird-eye view of a highway in Los Angeles."
67. "Campfire at night in a snowy forest with starry sky in the background."
68. "Cartoon character dancing."
69. "Cherry blossoms swing in front of ocean view, 4k, high resolution."
70. "Chocolate syrup pouring on a vanilla ice cream"
71. "Close up of grapes on a rotating table. High definition."
72. "Teddy bear surfer rides the wave in the tropics."
73. "Drone flythrough interior of sagrada familia cathedral."
74. "Drone flythrough of a fast food restaurant on a dystopian alien planet."
75. "Drone flythrough of a tropical jungle covered in snow."
76. "Edge lit hamster writing code on a computer in front of a scifi steampunk machine with yellow green red and blue gears high quality dslr professional photograph"
77. "Epic tracking shot of a Powerful silverback gorilla male walking gracefully",
78. "Fireworks."
79. "Flying through a temple in ruins, epic, forest overgrown, columns cinematic, detailed, atmospheric, epic, mist, photo-realistic, concept art, volumetric light, cinematic epic, 8k"
80. "Flying through an intense battle between pirate ships in a stormy ocean."
81. "Flying through fantasy landscapes, 4k, high resolution."
82. "Pug dog listening to music with big headphones."
83. "Horror house living room interior overview design, Moebius, Greg Rutkowski, Zabrocki, Karlkka, Jayison Devadas, Phuoc Quan, trending on Artstation, 8K, ultra wide angle, pincushion lens effect."
84. "Hulk wearing virtual reality goggles, 4k, high resolution."
85. "Incredibly detailed science fiction scene set on an alien planet, view of a marketplace. Pixel art."
86. "Ironman flying over a burning city, very detailed surroundings, cities are blazing, shiny iron man suit, realistic, 4k ultra high definition."
87. "Jack russell terrier dog snowboarding. GoPro shot."
88. "Low angle of pouring beer into a glass cup."
89. "Melting pistachio ice cream dripping down the cone."
90. "Milk dripping into a cup of coffee, high definition, 4k."
91. "Pouring latte art into a silver cup with a golden spoon next to it."
92. "Shoveling snow."
93. "Studio shot of minimal kinetic sculpture made from thin wire shaped like a bird on white background."

94. "Sunset time lapse at the beach with moving clouds and colors in the sky, 4k, high resolution."
95. "Teddy bear walking down 5th Avenue, front view, beautiful sunset, close up, high definition, 4k."
96. "The Orient Express driving through a fantasy landscape, animated oil on canvas."
97. "Time lapse at a fantasy landscape, 4k, high resolution."
98. "Time lapse at the snow land with aurora in the sky, 4k, high resolution."
99. "Tiny plant sprout coming out of the ground."
100. "Toy Poodle dog rides a penny board outdoors"
101. "Traveler walking alone in the misty forest at sunset."
102. "Turtle swimming in ocean."
103. "Two pandas discussing an academic paper."
104. "Two pandas sitting at a table playing cards, 4k, high resolution."
105. "Two raccoons are playing drum kit in NYC Times Square, 4k, high resolution."
106. "Two raccoons reading books in NYC Times Square."
107. "US flag waving on massive sunrise clouds in bad weather. High quality, dynamic, detailed, cinematic lighting"
108. "View of a castle with fantastically high towers reaching into the clouds in a hilly forest at dawn."
109. "Waves crashing against a lone lighthouse, ominous lighting."
110. "Woman in white dress waving on top of a mountain."
111. "Wood on fire."
112. "Wooden figurine surfing on a surfboard in space."
113. "Yellow flowers swing in wind."