

SketchVideo: Sketch-based Video Generation and Editing

Supplementary Material

Feng-Lin Liu^{1,2} Hongbo Fu³ Xintao Wang⁴ Weicai Ye⁴ Pengfei Wan⁴ Di Zhang⁴ Lin Gao^{1,2*}

¹Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences ³Hong Kong University of Science and Technology ⁴Kuaishou Technology

In this supplemental material, we provide additional implementation details (Sec. 1), showcase results of sketch-based video generation (Sec. 2) and editing (Sec. 3), and the results of consecutive generation and editing (Sec. 4). We also present additional ablation study (Sec. 5), memory and time analysis (Sec. 6), comparison results (Sec. 7), attention map visualization (Sec. 8), and failure cases (Sec. 9). We discuss ethical concerns in Sec. 10 and show the full text prompts of all examples in Sec. 11.

1. Implementation Details

Sketch-based Generation. During training, we use a learning rate of $1e-5$ and employ mixed precision (fp16). We optimize the sketch-conditioned network using the AdamW optimizer [10], while fixing the weights of the pretrained CogVideoX-2b [6] model. Similar to ControlNet [19], we initialize the weights of the DiT block with those from CogVideoX-2b. In the hybrid training stage (image and video), each batch consists solely of images or videos, with one or two keyframe sketches conditioned on random time points. The training dataset includes 53w video clips and 90w images, with the labeled text prompts from OpenVid [11] and LAION [13].

During inference, generated videos have a resolution of 720×480 and 8 fps, spanning 6 seconds with 49 frames. The classifier-free guidance scale is set to 10.0. Super-resolution [14] and frame interpolation [7] can be optionally applied to improve the resolution and frame rate, as in CogVideo. All the results shown in the paper and supplemental material do not use them.

Sketch-based Editing. We initialize the editing network using the weights from the sketch-based generation model. During training, the rectangle masks have random position, height, and weight. Mask movement strategies include: 1) fixed position, 2) linear interpolation between two random positions, and 3) motion based on the mean optical flow of the pixels within the box. The masks are directly multiplied with the input videos at the pixel level to remove the information in the edited regions.

During inference, the classifier-free guidance scale is set to 20.0. Users may manually adjust mask positions through interpolation or dynamic movement using optical flow. The masks are further downsampled and merged to achieve latent fusion. The 3D VAE in CogVideoX-2b performs temporally 8×8 spatial and $4 \times$ temporal downsampling. So,

we also downsample the masks by 8×8 to match the latent code resolution and combine the masks of four frames corresponding to one latent code. The masks are shown with orange boxes in the original videos in Fig. 7 and 8.

2. Sketch-based Generation Results

We show additional sketch-based video generation results to validate the effectiveness of our method. All results in this section are generated from **hand-drawn sketches** rather than synthetic ones. Fig. 1 shows the video generation results using a single keyframe sketch as input, conditioned on different time points (0s, 1.5s, 3s, 4.5s, and 6s). Our method generates realistic video results across diverse categories. Fig. 2 shows the results from the same sketch and text prompts, but with varying time points (0s, 1.5s, 3s, 4.5s, and 6s). These results exhibit fidelity to the sketch at the specified time point while introducing diverse motion in other frames.

In Fig. 3, we present additional video generation results from two keyframe sketches. Our method generates realistic videos with detailed control of geometry and motion. Fig. 4 shows the results with the same two keyframe sketches and text prompts, but varying time points. It can be seen that our method generates interesting interpolation and extrapolation results.

Our method uses the sketch to define geometry and the text prompt to specify appearance. As shown in Fig. 5, given the identical text but varying sketches, our method generates results of similar objects but with different geometry components and layouts. As shown in Fig. 6, our method generates diverse results from identical sketches with different text prompts, varying the appearance while maintaining the same structure.

3. Sketch-based Editing Results

We show additional results of sketch-base video editing in Fig. 7. Our method effectively edits object components within the video, such as modifying a person’s face or changing the roof of a castle. Our method also supports editing based on two keyframe sketches at different time points, controlling the motion of the newly generated objects (e.g., the transformation of a hat or the posing of a fox).

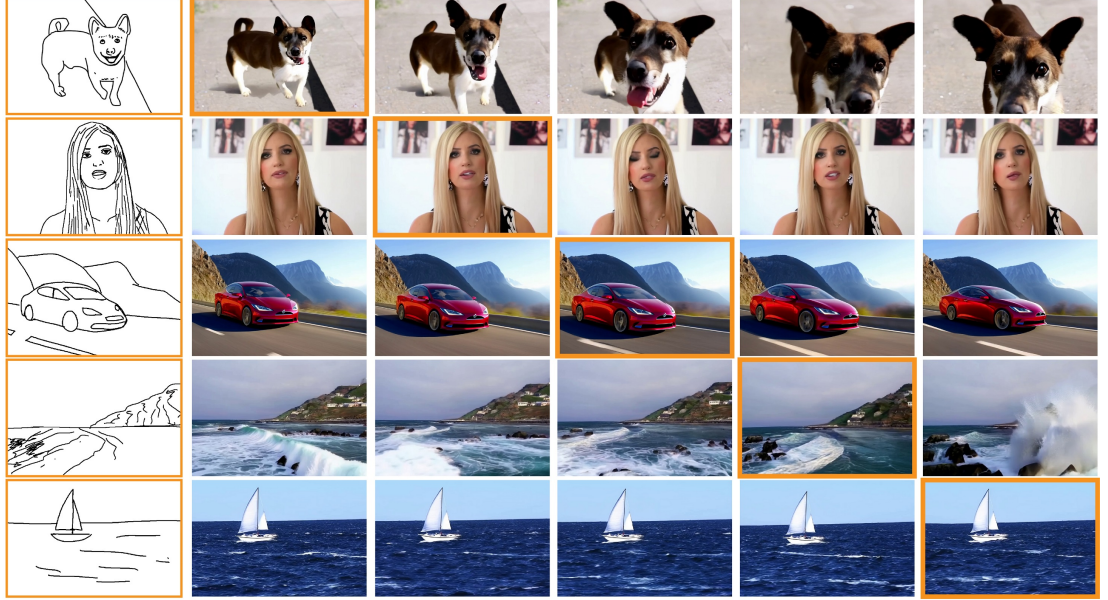
A very cute little wild dog with a mouth and shiny fur, shining brightly towards the center of the picture, very lively. The dog takes a walk on the street with a natural flow of movements.

The video features a woman with long blonde hair, wearing a black and white dress, and large earrings. She is seated in a room with a white wall, which is adorned with various photographs.

A red Tesla Model S drives on a well-maintained highway through a mountainous landscape at sunset, capturing the thrill of a high-performance electric vehicle in a stunning natural setting.

A video captures a serene coastal scene with a rocky shoreline, a hillside town, and crashing waves, blending nature and human habitation.

A high-angle view captures a white-sailed boat journeying across choppy waters on a sunny, windy day, moving rightward with a billowing sail.

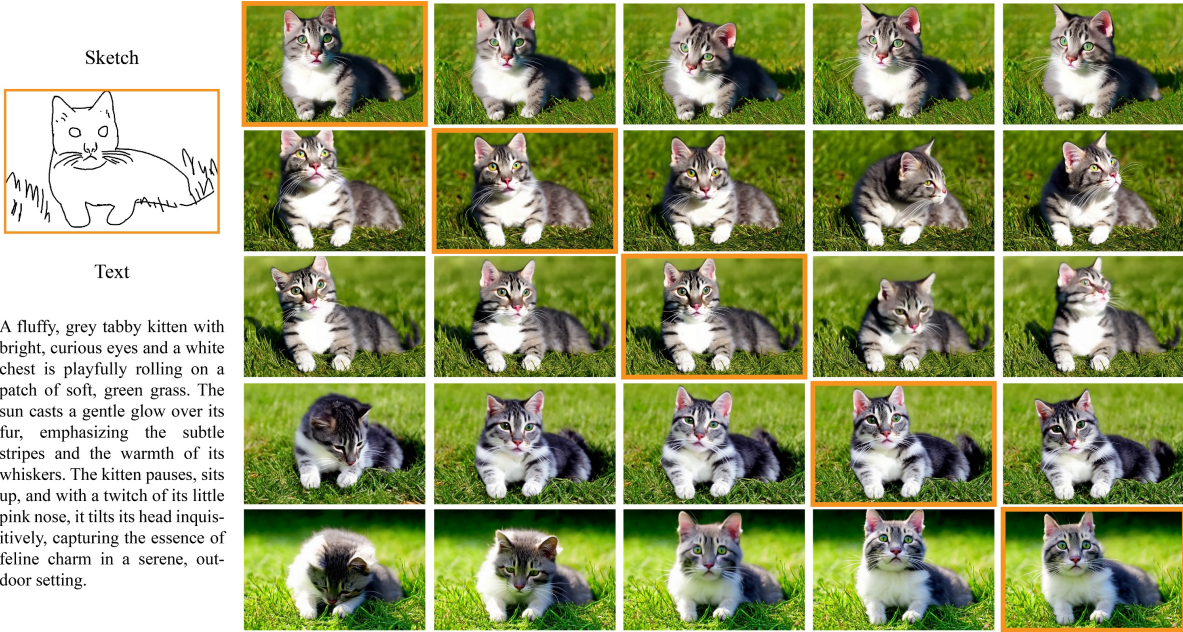


(a) Text

(b) Sketch

(c) Video Generation Results

Figure 1. Video generation results using one keyframe sketch. Given the text prompts (a) and sketches (b), our method generates realistic video results (c). The frames corresponding to the given time points are highlighted in orange boxes.



(a) Input

(b) Video Generation Results

Figure 2. Video generation results using the same sketch and text inputs (a) with different time points. The frames at 0s, 1.5s, 3s, 4.5s, and 6s (from top to bottom) are marked with orange boxes. The results (b) have good faithfulness to the sketch at given time points, while exhibiting diverse motion in other frames.

4. Consecutive Editing Results

Our method allows for consecutive generation and editing of videos. As shown in Fig. 8, given an initial sketch and a text prompt, we first generate a realistic video (Generation 1). The generated video then serves as input for further editing, using a new text prompt and a new sketch to modify

the content (Editing 1). Additional edits (Editing 2) can be made iteratively without degrading the quality of the generated video. This showcases the detailed control our method offers for video customization.

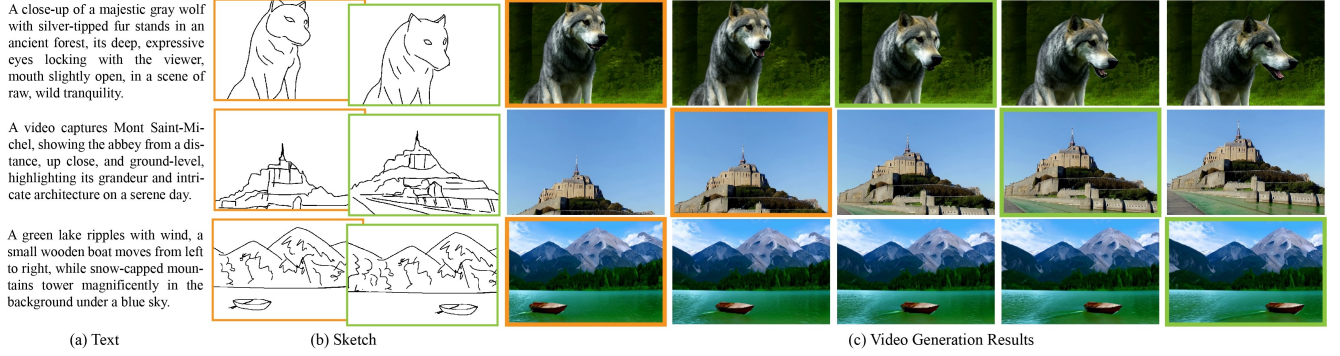


Figure 3. Video generation results using two keyframe sketches. Given the text prompts (a) and two sketches (b), our method generates realistic video results (c). The frames corresponding to the given time points are highlighted in orange and green boxes.



Figure 4. Video generation results (b) using the same two keyframe sketches and text inputs (a) at diverse time points. Each column in (b) represents frames at 0s, 1.5s, 3s, 4.5s, and 6s (from left to right). The frames corresponding to the given time points are highlighted in orange and green boxes.

5. More Ablation Study

During the editing process, we employ a latent fusion strategy to seamlessly fuse the newly generated regions with the original video content. An alternative approach involves directly fusing the generated video with the input video in pixel level. As shown in Fig. 9 (b), this baseline approach results in obvious seams along the boundaries of the edited regions. In contrast, our method ensures a smooth fusion of the edited and unedited regions by accurately inserting information from the unedited regions during the denoising process. And the combination in latent space followed with VAE decoding generates more realistic images than direct pixel level fusion.

Distribution of control blocks. We constructed a control branch with blocks at positions 1-5, 13-17, 26-30, and

Metric	1-5	13-17	26-30	Ours
LPIS↓	33.37	38.57	50.34	32.47

Table 1. Different distributions of sketch control blocks. Our method has the best sketch faithfulness compared with other distributions.

Method	Training			Inference			
	Ctrl-10	Ctrl-5	Ours	SpCtrl	Ctrl-10	Ctrl-5	Ours
Memory(GB)	70.79	55.37	56.84	11.94	21.23	19.95	20.35
Time(s)	N/A	N/A	N/A	29	52	46	41

Table 2. The memory and time efficiency of different methods. SpCtrl refers to SparseCtrl [5] method. Ctrl-10 and Ctrl-5 refer to the Ctrl-CogVideo baseline with 10 and 5 control blocks, respectively.



Figure 5. Video generation results using the same text prompts (a) and different keyframe sketches (b). The generated results (c) have a similar appearance while exhibiting different geometry details. The keyframe sketches and corresponding generated frames are highlighted in orange and green box.



Figure 6. The video generation results using the same two keyframe sketches (b) and different text prompts (a). The generated results (c) share a similar structure while exhibiting different appearances, demonstrating the influence of text prompts on visual details. The keyframe sketches and corresponding generated frames are highlighted in the orange and green boxes.

a uniform distribution (ours). To evaluate the effectiveness of these placements, we compute LPIPS between the input and extracted sketches from corresponding generated frames. Due to the restriction of computing resources, we trained 10,000 steps on our hybrid image and video dataset. As shown in Table 1, our method incorporates control signals at multiple feature levels, enabling a more comprehensive analysis and achieving superior performance compared to the baselines. Additionally, in the DiT-based video diffusion model (CogVideo-2b in our method), we observe that earlier blocks primarily influence geometric features, while later blocks exhibit reduced controllability in this aspect.

6. Memory and Time Efficiency

A direct comparison of memory and time with existing methods, such as SparseCtrl, is not meaningful due to different architectures (Unet for SparseCtrl, DiT for ours) and frame counts (16 vs. 49 frames). We applied SparseCtrl’s concept to CogVideo, but using half of the blocks (15 blocks) as in Pixart- δ [1] resulted in out-of-memory errors. In Table 2, we report its memory and time with 10 blocks (Ctrl-10) and 5 blocks (Ctrl-5) below. Our sketch control block consumes slightly more memory compared with Ctrl-5 due to additional inter-frame attention. Our method achieves faster inference speed compared with the Ctrl-5 and Ctrl-10 baselines.

7. More Comparison Results

We present additional comparison results for both generation and editing. For sketch-based generation, Fig. 11

shows comparison results of a single keyframe sketch. We exclude results from video interpolation methods [2, 9, 15] since they require two keyframes. Our method generates results that are most faithful to the input sketches, such as the wing shape and a notch in the cake. Fig. 12 shows comparison results for two keyframe sketches. We compare with two additional video interpolation methods, including Seine [2] and ToonCrafter [15], whose official codes and weights are utilized in our experiments. Since the text-to-image results of two keyframes are not consistent, these methods generate fuzzy details in intermediate frames. Our method generates the most realistic results with the best temporal coherence.

For sketch-based editing, Fig. 13 shows additional comparison results, including two additional methods: TokenFlow [4] and I2VEdit [12]. Our method generates the most realistic and geometrically consistent results. In contrast, TokenFlow [4] exhibits minor editing effects due to the complexity of geometry modifications, while InsV2V [3] has a similar problem with insufficient geometry control. Furthermore, AnyV2V [8] and I2VEdit [12] totally change the unedited regions and have fuzzy details.

Quantitative Comparison. We conducted quantitative experiments, including additional generation methods (i.e., Seine [2] and ToonCrafter [15]) and editing methods (i.e., TokenFlow [4] and I2VEdit [12]). For generations, Seine and ToonCrafter achieved similar sketch faithfulness to AMT, as they utilize the same beginning and ending frames. However, they exhibit lower temporal consistency due to the smaller number of frames and greater content variation between adjacent frames. For editing, existing methods sig-



Figure 7. Results of sketch-based video editing. For each example, the input text and drawing sketch(es) are given in (a). In (b), the original video is at the top, while the edited video is at the bottom. The sketches and corresponding frames are highlighted in orange and green boxes.

Metrics	ToonCrafter[15]	Seine[2]	AMT[9]	SparseCtrl[5]	Ctrl-CogVideo	Ours
LPIPS ↓	29.09	29.54	29.17	44.85	32.23	27.56
CLIP ↑	95.75	92.76	96.12	96.48	98.04	98.31

Table 3. The quantitative results of sketch-based video generation comparison. The LPIPS and CLIP numbers are scaled up 100×, with each cell colored to indicate the **best**. Methods ToonCrafter [15], Seine [2], and AMT [9] interpolate ControlNet-generated images while the other method directly translate sketches into videos.

nificantly change the unedited regions, as reflected in their low reconstruction values in these regions. Our method has the best sketch faithfulness, temporal consistency, and unedited region preservation.

8. Sketch Propagation Illustration

As shown in Fig. 10, we visualize the attention maps during the generation of a frame highlighted in an orange box, both for input sketches (b, c) and itself (d). They are consistent

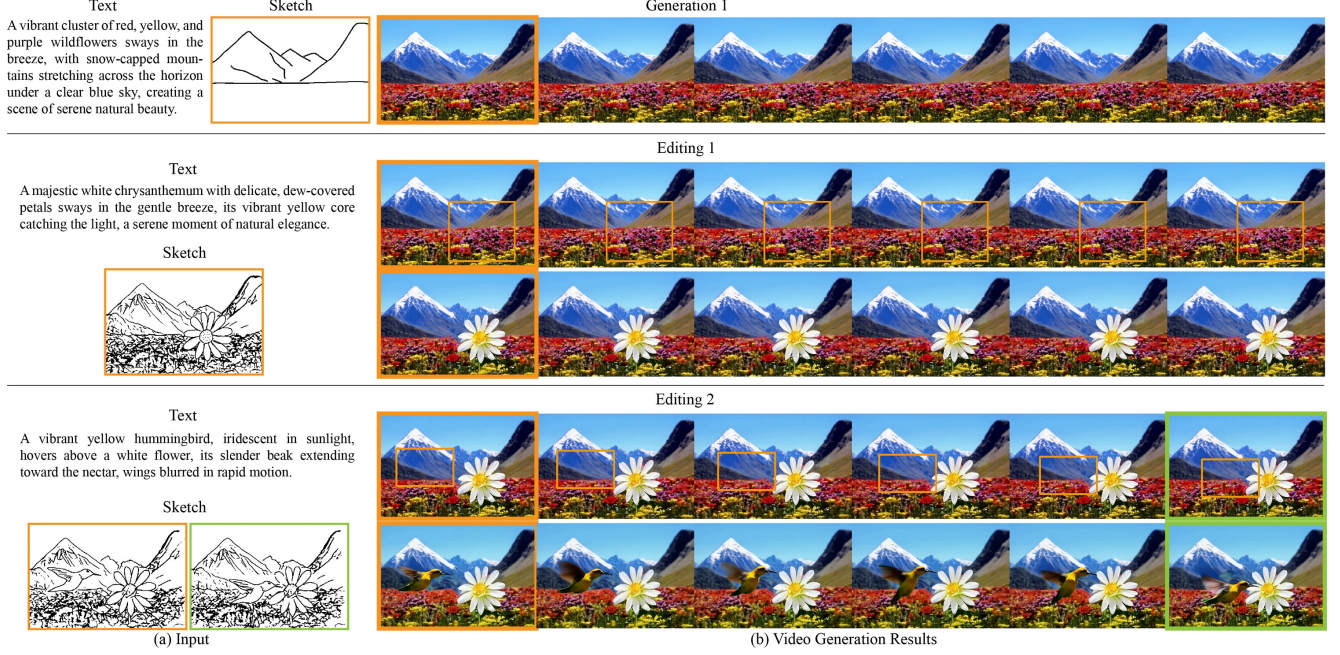


Figure 8. Consecutive video generation and editing results. In Step 1 (Generation 1), a video is generated based on an input text and a sketch. In Step 2 (Editing 1), a flower is added to the generated video. Step 3 (Editing 2) adds a hummingbird to the edited video.

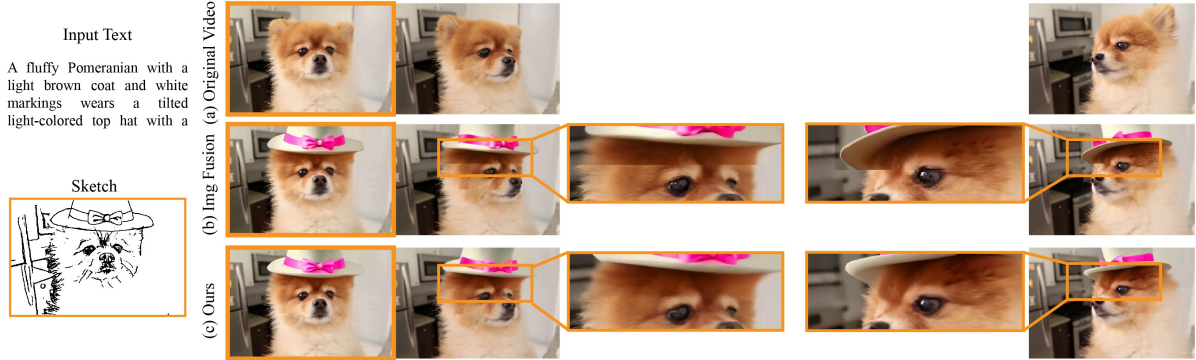


Figure 9. Ablation study of latent fusion. Replacing the latent fusion strategy with direct image space fusion leads to edited results (b) with noticeable seams. Our method removes this artifact and generates a realistic edited video with the new component added into the original content.

with the input sketches and frame’s edge maps, supporting our inter-frame attention mechanism for edge analysis and sketch-based controllability.

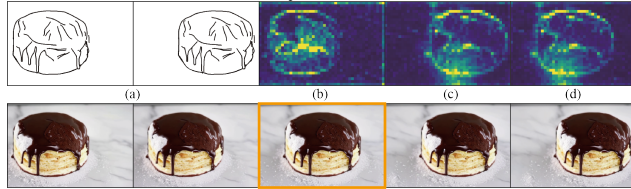


Figure 10. The visualization of inter-frame attention maps. The input sketches (a) and attention maps (b,c,d) are shown in the top row, while the generated video is shown in the bottom row.

9. Failure Cases

Fig. 14 shows failure cases of our method. Similar to image generation, our method generates fuzzy details in challenging cases, such as the human hands. Additionally, when the two keyframe sketches differ significantly in content, the generated video transitions may appear unnatural or incoherent due to the lack of such a dataset. 3D geometry information and data augmentation can be potentially used to solve these problems.

10. Ethical Discussion

Our method is designed for positive applications such as movie and short video production. We strongly condemn the abuse of AI video generation technologies. The meth-

Metrics	InsV2V[3]	AnyV2V[8]	I2VEdit[12]	TokenFlow[4]	Ours
LPIPS ↓	13.61	11.92	10.98	12.92	9.74
CLIP ↑	95.39	93.47	96.81	97.83	98.34
PSNR ↑	16.84	13.68	12.98	21.15	36.48

Table 4. The quantitative results of sketch-based video editing comparison. The LPIPS and CLIP numbers are scaled up 100×, with each cell colored to indicate the **best**. AnyV2V [8] and I2VEdit [12] propagate the image editing results into videos. InV2V [3] and TokenFlow [4] utilize text prompts for video editing.



Figure 11. Comparison of sketch-based video generation results for one keyframe sketch input. We compare our method with SparseCtrl [5] and Ctrl-CogVideo (baseline discussion in Sec 4.3 in main paper). Our method generates the results that are most faithful to the input sketches.

ods [16–18] that detect the fake video may be helpful to avoid the potential misuse of existing video generation approaches. Our method can also be used to generate training data for these detection methods.

11. Full Text Prompt

For ease of reading, the text prompts shown in the figures are simplified. The full-text prompts of each figure are shown in the following:

Main Paper

Figure 1:

The camera meticulously frames a close-up of a majestic male lion’s face, his mane a regal cascade of tawny hues, eyes a piercing amber gaze that speaks of the wild spirit within. The sunlight bathes his features in a golden glow, highlighting the rugged texture of his fur and the subtle scars that tell tales of past battles. His whiskers are pro-

nounced, sensitive to the faintest breeze, and his mouth is slightly agape, revealing formidable teeth, a silent testament to his power as the king of the savannah.

The video is a drone shot of a serene landscape featuring a deep blue lagoon surrounded by lush green trees and rocky cliffs. The lagoon is nestled between the mountains, creating a picturesque scene. The drone captures the tranquility of the water and the vibrant colors of the surrounding nature. The video is a beautiful representation of the natural beauty of the area, showcasing the lagoon’s unique location and the stunning views it offers.

The video captures the breathtaking beauty of a mountainous landscape. The first frame shows a clear blue sky above the mountains, with a few clouds scattered across it. The second frame reveals a deep blue lake nestled between the mountains, its calm waters reflecting the surrounding scenery. The third frame offers a closer view of the mountains, showcasing their rugged terrain and the patches of

A large, gray teddy bear with a blue ribbon holds a small, white cake with a single candle, sitting in a brightly lit room with a white wall, eyes closed, suggesting a birthday celebration.

Input Text



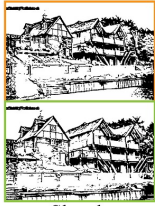
Sketch



Text-to-Image Results

A serene waterfront scene with two large, stone and wood houses on a grassy hillside, a wooden dock with a small boat, and calm water reflecting the tranquil, lush surroundings.

Input Text



Sketch



Text-to-Image Results

(a) AMT

(b) Seine

(c) ToonCrafter

(d) SparseCtrl

(e) Ctrl-CogVideo

(f) Ours

(a) AMT

(b) Seine

(c) ToonCrafter

(d) SparseCtrl

(e) Ctrl-CogVideo

(f) Ours



Figure 12. Comparison of sketch-based video generation results for two keyframe sketches. The input text prompts and sketches are shown in the left region, while the results of AMT [9], Seine [2], ToonCrafter [15], SparseCtrl [5], Ctrl-CogVideo (baseline discussed in Sec 4.3 in the main paper), and our method are shown in the right region. Our method achieves the best temporal coherence in the generated video.

snow that still cling to their peaks. The overall style of the video is serene and majestic, capturing the natural beauty of the landscape in a way that is both awe-inspiring and tranquil.

A magnificent waterfall cascades down from a series of rocky cliffs. The waterfall occupies most of the area in the picture, which is very magnificent and spectacular. The water flow of the waterfall is very turbulent, with white splashes spanning the entire scene. The waterfall cascades

from the top of the cliff all the way to the bottom of the lake.

A fluffy Pomeranian dog wears a light-colored top hat made of cloth adorned with a pink satin bow tied neatly around the base of the hat. The dog has a light brown coat with white markings on its face and chest. The top hat, slightly tilted, gives the Pomeranian a charming and playful appearance, while the pink bow adds a touch of elegance and whimsy. The dog’s fur is well-groomed, showcasing its signature soft, voluminous coat that frames its small face.

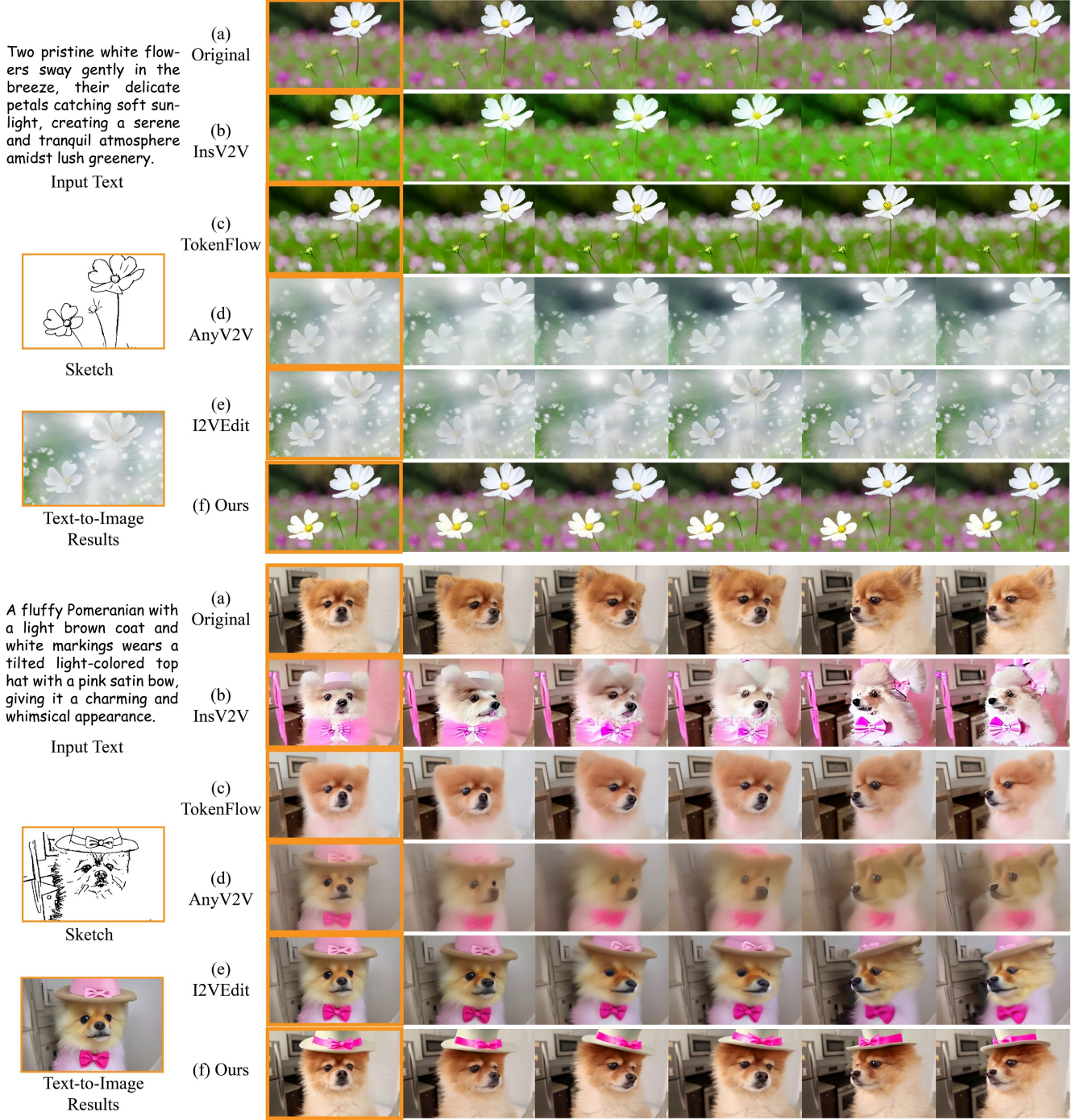


Figure 13. Comparison of video editing results for different methods. The input text prompts and sketches are shown on the left. The edited results from InsV2V [3], TokenFlow [4], AnyV2V [8], I2VEdit [12], and ours are shown on the right. Our method generates the most realistic video editing results.

Figure 3:

A small, fluffy rabbit with soft, white fur and a pink nose hops gracefully from left to right across a lush, green meadow. The sun is setting, casting a warm, golden glow over the scene. The rabbit pauses momentarily, its ears twitching as it listens to the gentle rustling of the leaves in

the nearby trees. It then continues its playful dance, moving side to side with an air of innocence and joy, the grass beneath its paws a soft carpet of nature’s embrace.

A sweeping panoramic video captures the grandeur of a majestic city, where towering skyscrapers pierce the clouds, their glass facades shimmering in the sunlight. Nestled at

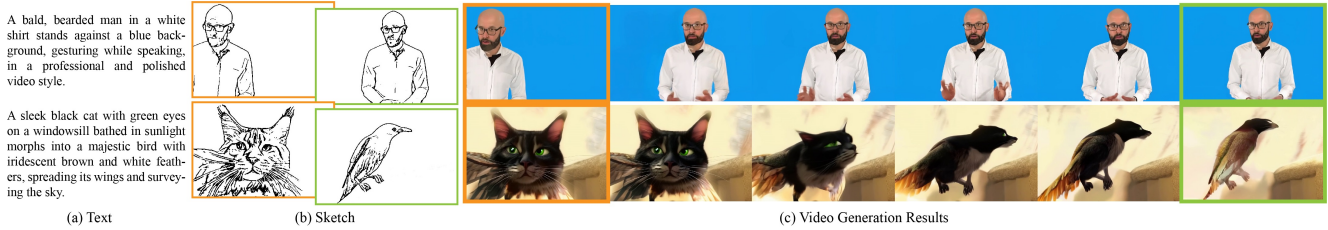


Figure 14. Failure cases of our method. Our method generates fuzzy details in hands and eyeglasses. Performance degrades when the two keyframe sketches (highlighted in orange and green box) depict entirely different objects, resulting in strange transitions between frames.

the base of these architectural giants, quaint urban homes with red-tiled roofs and leafy gardens form a charming tapestry around the city’s perimeter. In the bottom left corner, a tranquil lake mirrors the sky with its rippling waves of azure water, while fluffy white clouds drift lazily across the expansive blue canvas above, completing the serene yet powerful urban landscape.

The video captures a close-up of a cat’s face in three frames. The cat has striking green eyes and a brown and black spotted coat. In the first frame, the cat is looking directly at the camera, its eyes wide and alert. In the second frame, the cat’s gaze is slightly averted, and its eyes are more relaxed. In the third frame, the cat is looking away from the camera, its eyes focused on something in the distance. The style of the video is a simple, yet intimate portrait of the cat, with a focus on its expressive eyes and fur pattern.

In a breathtaking cosmic event, a video captures the catastrophic explosion of a light blue planet, its surface fragmenting as rocks are catapulted into the void. The sky behind is a tapestry of stars, while the planet itself becomes a maelstrom of debris, with towering plumes of smoke and cascades of fiery orange flames painting a harrowing picture of celestial destruction. The detritus from the disintegrating planet creates a mesmerizing display against the serene backdrop of space, a stark contrast between the calm universe and the violent upheaval of the once-stable world.

The video is a time-lapse aerial shot of a large, ornate building with a distinctive dome and multiple balconies. The building is surrounded by lush greenery and a well-manicured lawn. In the first frame, the building is bathed in daylight, with the sun shining brightly on its facade. In the second frame, the sun begins to set, casting a warm glow on the building and its surroundings. In the third frame, the sun has set, and the building is illuminated by artificial lights, creating a stark contrast with the darkening sky. The style of the video is realistic, with a focus on the architectural details of the building and the natural beauty of its surroundings.

Figure 4:

Plumes of steam begin to billow from the volcano’s crest, hinting at the raw power simmering beneath the surface, a prelude to the impending eruption that promises to reshape the landscape. Rising majestically in the distance, a colos-

sal volcano stands as a silent sentinel, its peak dusted with a delicate blanket of white snow. The volcano’s flanks are lush with verdant vegetation, a vibrant green tapestry that contrasts starkly against the earthy tones of its rocky surface. In the foreground of the screen, the lake is rippling with water waves

A man is speaking. He has thick hair and is wearing a black suit with a white shirt and a red tie underneath. The camera focuses on his face and upper body.

The ornamental fish swims from left to right. A mesmerizing ornamental fish glides through the inky depths of the ocean, its vibrant scales shimmering with hues of electric blue, fiery orange, and iridescent green. With a sleek, flattened body that navigates the vast underwater expanse. Its delicate fins, adorned with intricate patterns, flutter and sway like rhythmic poetry in motion, leaving a trail of iridescence in the silent sea.

A green grassland and a small path extending into the distance. The grassland is covered with gravel.

Figure 5:

The video captures a large cruise ship docked at a pier. The ship is painted in white and black, with a distinctive yellow and gold design on its side. The ship is moored to the pier, which is lined with several smaller boats. The sky above is filled with clouds, suggesting an overcast day. The water around the ship is calm, reflecting the ship’s grandeur. The pier extends into the water, providing a clear view of the ship’s size and scale. The video is a time-lapse, capturing the ship’s arrival and docking process. The style of the video is realistic, with a focus on the ship and its surroundings. The video does not contain any people or animals, and the focus is solely on the ship and its immediate environment.

Figure 6:

A vibrant magpie, its feathers a striking contrast of soft white and cool black, perches gracefully atop a slender branch of an ancient flower tree. The bird’s delicate talons grip the bark with precision, as soft morning light filters through the leafless canopy, casting gentle shadows that dance around it. The magpie’s head tilts slightly, its keen eye surveying the tranquil winter scene, while the rest of the forest lies quiet and still in the crisp, cool air of an early dawn.

Figure 7:

Extreme close-up of chicken and green pepper kebabs grilling on a barbecue with flames. Shallow focus and light smoke. vivid colours

Figure 8:

The video captures a wooden boat with coastal landscape. A small wooden boat is sailing on the sea surface. The wooden boat is composed of wooden boards with rich textures, with a rustic texture.

Supplemental Material

Figure 1:

A very cute little wild dog with a mouth and shiny fur, shining brightly towards the center of the picture, very lively. The dog takes a walk on the street with a natural flow of movements.

The video features a woman with long blonde hair, wearing a black and white dress, and large earrings. She is seated in a room with a white wall, which is adorned with various photographs. The woman is looking directly at the camera, and her expression is neutral. The room has a bright and airy feel to it. The style of the video is a combination of a personal interview and a motivational message.

The video features a red Tesla Model S driving on a highway with a mountainous landscape in the background. The car is sleek and modern, with a streamlined design and a large front grille. The highway is wide and well-maintained, with clear lane markings. The mountains rise majestically in the distance, their peaks shrouded in mist. The sky is a clear blue, and the sun is setting, casting a warm glow over the scene. The car is moving at a steady pace, and the driver appears to be focused on the road ahead. The overall style of the video is dynamic and energetic, capturing the thrill of driving a high-performance electric vehicle in a beautiful natural setting.

The video captures a serene coastal scene, showcasing the natural beauty of the area. The first frame shows a rocky shoreline with the calm ocean gently lapping against the rocks. The second frame reveals a small town nestled on the hillside, with houses and buildings scattered across the landscape. The third frame offers a closer view of the shoreline, with the waves crashing against the rocks, creating a dynamic contrast between the stillness of the ocean and the power of the waves. The overall style of the video is a blend of nature and human habitation, with a focus on the interplay between the land and the sea.

The video captures a sailboat journey across a vast body of water. The sailboat, with its white sail billowing in the wind, is the main focus of the video. The boat is seen from a high angle, providing a bird's eye view of its journey. The water around the boat is choppy, indicating a windy day. The sky above is a clear blue, suggesting a sunny day. The boat is moving towards the right side of the frame, indicating its direction of travel. The overall style of the video is a dynamic and adventurous depiction of a sailboat journey

on a windy day.

Figure 2:

A fluffy, grey tabby kitten with bright, curious eyes and a white chest is playfully rolling on a patch of soft, green grass. The sun casts a gentle glow over its fur, emphasizing the subtle stripes and the warmth of its whiskers. The kitten pauses, sits up, and with a twitch of its little pink nose, it tilts its head inquisitively, capturing the essence of feline charm in a serene, outdoor setting.

Figure 3:

In a breathtaking close-up, a majestic gray wolf with a coat of realistic, silver-tipped fur stands regally against the lush, green tapestry of an ancient forest. Its eyes, deep and expressive, seem to hold the secrets of the wild, locking gazes with the viewer through the lens. With a mouth agape, the wolf appears to be caught in a silent howl or a gentle pant, breathing life into the serene scene. The atmosphere is one of unspoken grandeur, as the video encapsulates the raw, wild essence of the wolf amidst the tranquility of its untouched habitat.

The video captures the majestic Mont Saint-Michel, a historic abbey and fortification located on a small island in Normandy, France. The first frame shows the abbey from a distance, its grandeur accentuated by the clear blue sky. The second frame offers a closer view, revealing the intricate details of the abbey's architecture. The third frame provides a ground-level perspective, allowing viewers to appreciate the scale of the abbey and its surroundings. The video is a testament to the abbey's historical significance and architectural beauty, set against the backdrop of a serene day.

At the bottom of the screen is a green lake surface, rippling with the wind. A small wooden boat moved from the left side of the lake to the right side. In the middle are continuous mountains, towering and magnificent, not satisfied with the green forest, with snow mixed at the top of the mountains. The background is a blue sky.

Figure 4:

A bustling scene at the Piazza del Duomo in Milan, Italy. The focal point of the scene is the large archway entrance to the Galleria Vittorio Emanuele II, a historic shopping mall. The archway, constructed of white stone, stands majestically against the backdrop of a clear blue sky. Flanking the entrance are two buildings, one on each side. These buildings, also made of white stone, are adorned with green awnings, adding a touch of color to the otherwise monochrome structure. The piazza itself is teeming with life. People can be seen walking around, adding a dynamic element to the scene. In the background, a few trees can be spotted, providing a touch of nature amidst the urban setting. Despite the scene being taken from a distance, the details of the archway and the surrounding buildings are clearly visible, showcasing the architectural grandeur of the landmark. The scene does not contain any discernible text.

The relative positions of the objects suggest a well-planned architectural design, with the archway centrally located and the buildings symmetrically placed on either side.

Figure 5:

The video captures a brown owl in flight, soaring through a forested area. The owl's wings are spread wide, showcasing its impressive wingspan. The owl's eyes are wide open, alert and focused on its surroundings. The background is a blur of green and brown, indicating the presence of trees and foliage. The owl's flight path is smooth and graceful, demonstrating the bird's agility and control. The overall style of the video is a close-up, slow-motion shot, allowing viewers to appreciate the details of the owl's flight and the beauty of its natural habitat.

Figure 6:

The video shows a chocolate cake being decorated with candy. The cake is placed on a white plate, and the candies are scattered on top of the cake. The candies are orange and yellow, and they are placed on top of the chocolate frosting. The cake is then drizzled with chocolate sauce, which is poured over the top of the cake. The sauce is thick and glossy, and it drips down the sides of the cake. The cake is then placed on a table, and the camera zooms in on the cake, showing the details of the decoration. The video is a close-up shot, focusing on the cake and the decoration. The style of the video is a food video, showcasing the process of decorating a cake.

A gleaming yellow cake, adorned with white frosting and delicate piping, sits at the center of a marble countertop. Above it, a cascade of vibrant pink strawberry juice, shimmering under the kitchen lights, gently pours from a height, drizzling down the sides of the cake, pooling slightly around the base, and creating a mesmerizing contrast between the yellow cake and the rosy liquid.

Figure 7:

The video shows a man with a beard standing next to a black truck. He is wearing a black T-shirt. In the first frame, he is pointing at the truck. In the second frame, he is opening the door of the truck. In the third frame, he is sitting inside the truck. The truck is parked in a lot with trees in the background. The man appears to be in the process of getting into the truck. The style of the video is casual and informal.

The video captures the majestic Neuschwanstein Castle, a 19th-century Romanesque Revival palace, perched on a rugged hill above the village of Hohenschwangau near Füssen in southwest Bavaria, Germany. The castle, a symbol of fairy tales and fantasy, is seen in three different frames, each showcasing its grandeur against the backdrop of the snow-capped mountains and the surrounding forest. The first frame offers a distant view of the castle, its multiple towers and turrets reaching towards the sky. The second frame provides a closer perspective, revealing the intricate

details of the castle's architecture. The third frame offers a panoramic view of the castle, its towers and turrets standing tall against the backdrop of the snow-capped mountains and the surrounding forest. The video is a testament to the castle's historical significance and architectural beauty.

A close-up shot captures the innocent, yet adventurous expression of a young, blonde girl, her eyes a shimmering shade of blue. She's adorned in a classic khaki canvas top hat, casting a gentle shadow over her bright, curious eyes. The scene is one of quiet wonder, with the girl's face conveying a mix of mischief and wonder, as if she's about to embark on a grand, unknown journey.

An ancient beige white temple, crafted from marble with a subtle yellow patina, stands majestically amidst a serene landscape. It boasts an array of towering pillars, each meticulously carved, supporting a grand, beige white roof. The temple's weathered surface reveals the whispers of time, marked by intricate cracks and the subtle discolorations that speak to its storied past.

A small, light white fox with a fluffy tail and alert ears sits gracefully atop a lush green meadow, its head initially facing the viewer with a curious gaze. Gradually, the little fox tilts its head to the left, its eyes glinting with a mix of curiosity and wariness, as it surveys the surroundings. The soft sunlight filters through the nearby trees, casting dappled shadows that dance across the grass and the fox's fur, creating a tranquil and picturesque scene. The grassland in the background presents a bright yellow green color.

Figure 8:

A vibrant cluster of wildflowers, a kaleidoscope of reds, yellows, and purples, captures the foreground, their delicate petals swaying gently in the breeze. Beyond, a majestic range of snow-capped mountains stretches across the horizon, their peaks glistening under the midday sun, creating a breathtaking contrast against the clear blue sky. The flowers, a testament to the valley's vibrant life, and the distant mountains, symbols of enduring strength, together paint a picture of serene natural beauty.

A majestic white chrysanthemum, its circular petals meticulously wrapped around a vibrant yellow core, sways gracefully in a gentle breeze. Each petal, delicate and perfectly formed, shimmers with the morning dew, catching the light as the flower dances with the rhythm of the wind. The scene captures the delicate balance of nature, the floral beauty moving in harmony with the unseen forces around it, a serene moment of natural elegance.

A vibrant yellow hummingbird, its iridescent feathers shimmering in the soft sunlight, hovers with exquisite precision above a pristine white flower. The scene is captured in side profile, showcasing the bird's slender beak as it extends towards the flower's nectar-rich center, wings blurred by the rapidity of their motion, a mesmerizing dance of nature's delicate balance.

Figure 9:

A fluffy Pomeranian dog wears a light-colored top hat made of cloth adorned with a pink satin bow tied neatly around the base of the hat. The dog has a light brown coat with white markings on its face and chest. The top hat, slightly tilted, gives the Pomeranian a charming and playful appearance, while the pink bow adds a touch of elegance and whimsy. The dog's fur is well-groomed, showcasing its signature soft, voluminous coat that frames its small face.

Figure 10: A mouthwatering round cake, adorned with a dusting of powdered sugar, and a generous drizzle of glossy, dark chocolate ganache. The cake itself is a masterpiece of delicate sponge layers, each separated by a rich, velvety buttercream frosting that peeks through the sides, inviting everyone to take a slice. The scene is set with a soft focus on the cake, positioned on a marble countertop, with gentle lighting casting subtle shadows to accentuate its delectable texture.

Figure 11:

The video captures a brown owl in flight, soaring through a forested area. The owl's wings are spread wide, showcasing its impressive wingspan. The owl's eyes are wide open, alert and focused on its surroundings. The background is a blur of green and brown, indicating the presence of trees and foliage. The owl's flight path is smooth and graceful, demonstrating the bird's agility and control. The overall style of the video is a close-up, slow-motion shot, allowing viewers to appreciate the details of the owl's flight and the beauty of its natural habitat.

The video shows a close-up of three chocolate-covered desserts with various toppings, including nuts and a red filling, on a white plate. The desserts are presented in a way that suggests they are being eaten or sampled, with one dessert partially cut into and a spoon nearby. The style of the video is simple and straightforward, focusing on the desserts without any additional context or background. The lighting is bright, highlighting the textures and colors of the desserts. The video is likely intended to showcase the desserts' presentation and appeal to viewers' appetites.

Figure 12:

The video features a large, gray teddy bear with a white nose and a blue ribbon around its neck. The bear is holding a small cake with a single candle on it. The bear is sitting in a room with a white wall in the background. The bear appears to be looking at the cake. The video is likely a celebration of a birthday or a special occasion. The bear's fur is soft and fluffy, and the blue ribbon adds a touch of color to the gray bear. The cake is small and white, with a single candle on top. The bear's eyes are closed, and it seems to be enjoying the moment. The room is brightly lit, and the white wall in the background provides a clean and simple backdrop for the scene. The bear is the main focus of the video, and its size and position in the frame make it the

most prominent object. The cake is smaller in comparison, but it still stands out due to its bright color and the candle on top. The bear's position relative to the cake suggests that it is about to blow out the candle. The overall style of the video is simple and straightforward, with a focus on the bear and the cake. The lighting is bright and even, and the colors are soft and muted. The video does not contain any text or additional objects, and the focus is solely on the bear and the cake.

The video shows a serene waterfront scene with two large, multi-story houses situated on a grassy hillside. The houses are constructed with a combination of stone and wood, featuring traditional architectural elements such as pitched roofs and bay windows. The houses are surrounded by lush greenery, including a variety of trees and shrubs, which add to the natural beauty of the setting. In the foreground, there is a wooden dock extending into the water, with a small boat moored at the end. The boat appears to be a leisure vessel, possibly used for fishing or recreational purposes. The calm water reflects the houses and the surrounding landscape, creating a peaceful and idyllic atmosphere. The style of the video is realistic and naturalistic, capturing the tranquility of the waterfront location with a focus on the architecture and the natural environment. The camera angles and movements are steady and unhurried, allowing the viewer to appreciate the details of the houses and the surrounding landscape. The lighting is soft and diffused, suggesting either an overcast day or a time when the sun is not directly shining on the scene. Overall, the video presents a picturesque and serene waterfront setting, likely intended to evoke a sense of relaxation and escape from the hustle and bustle of everyday life.

Figure 13:

Two pristine white flowers gently sway in the gentle breeze, their delicate petals catching the soft rays of sunlight that filter through the surrounding foliage. Each flower, with its subtle fragrance and intricate design, appears almost weightless as it dances gracefully in the air, creating a serene and tranquil atmosphere. The movement of these blossoms against the backdrop of lush greenery adds a dynamic element to the otherwise still scene, evoking a sense of peace and beauty in nature.

A fluffy Pomeranian dog wears a light-colored top hat made of cloth adorned with a pink satin bow tied neatly around the base of the hat. The dog has a light brown coat with white markings on its face and chest. The top hat, slightly tilted, gives the Pomeranian a charming and playful appearance, while the pink bow adds a touch of elegance and whimsy. The dog's fur is well-groomed, showcasing its signature soft, voluminous coat that frames its small face.

Figure 14:

The video features a bald man with a beard and glasses, wearing a white shirt. He is standing against a blue back-

ground. The man appears to be speaking or presenting, as he is gesturing with his hands. The style of the video is professional and polished, suggesting it could be a corporate or educational video. The man's attire and the background give the impression of a formal or professional setting.

In a whimsical transformation, a sleek black cat with piercing green eyes sits gracefully on a windowsill, bathed in the warm glow of the afternoon sun. As the scene progresses, the cat blinks slowly, and with a shimmering aura, begins to morph seamlessly into a majestic bird, its fur transitioning into feathers of iridescent brown and white. The newly formed bird spreads its powerful wings, perched confidently on the same sill, now surveying the open sky with an air of freedom and wild spirit.

References

- [1] Junsong Chen, Yue Wu, Simian Luo, Enze Xie, Sayak Paul, Ping Luo, Hang Zhao, and Zhenguo Li. Pixart- δ : Fast and controllable image generation with latent consistency models. *CoRR*, abs/2401.05252, 2024. 4
- [2] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *Int. Conf. Learn. Represent.*, 2023. 4, 5, 8
- [3] Jiaxin Cheng, Tianjun Xiao, and Tong He. Consistent video-to-video transfer using synthetic dataset. In *Int. Conf. Learn. Represent.*, 2024. 4, 7, 9
- [4] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *Int. Conf. Learn. Represent.*, 2024. 4, 7, 9
- [5] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *Eur. Conf. Comput. Vis.*, pages 330–348, 2024. 3, 5, 7, 8
- [6] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1
- [7] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Eur. Conf. Comput. Vis.*, 2022. 1
- [8] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhui Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 4, 7, 9
- [9] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 4, 5, 8
- [10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, 2019. 1
- [11] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *CoRR*, abs/2407.02371, 2024. 1
- [12] Wenqi Ouyang, Yi Dong, Lei Yang, Jianlou Si, and Xingang Pan. I2vedit: First-frame-guided video editing via image-to-video diffusion models. *CoRR*, abs/2405.16537, 2024. 4, 7, 9
- [13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Adv. Neural Inform. Process. Syst.*, 2022. 1
- [14] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, pages 1905–1914, 2021. 1
- [15] Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncraft: Generative cartoon interpolation. *CoRR*, abs/2405.17933, 2024. 4, 5, 8
- [16] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Int. Conf. Comput. Vis.*, pages 22412–22423, 2023. 7
- [17] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Adv. Neural Inform. Process. Syst.*, pages 4534–4565, 2023.
- [18] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 7
- [19] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3813–3824, 2023. 1