Supplementary Material for DropletVideo: A Dataset and Approach to Explore Integral Spatio-Temporal Consistent Video Generation

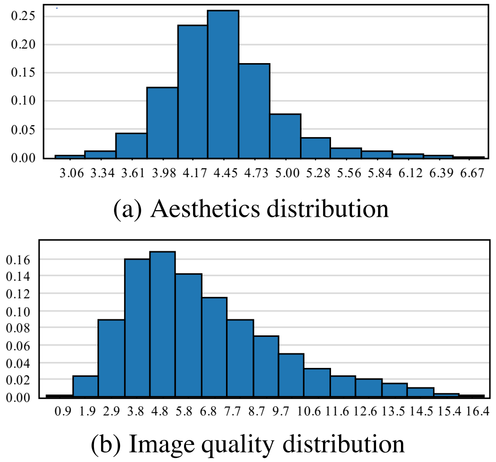
Runze Zhang[1,*], Guoguang Du[1,*], Xiaochuan Li[1,*], Qi Jia[1,*], Liang Jin[1,*], Lu Liu[1], Jingjing Wang[1]
Cong Xu[1], Zhenhua Guo[1], Yaqian Zhao[1], Xiaoli Gong[2], Rengang Li[3,1,†], Baoyu Fan[2,1,†]
[1]IEIT SYSTEMS Co., Ltd. [2]Nankai University [3]Tsinghua University
Jinan, China Tianjin, China Beijing, China
{zhangrunze, duguoguang, lixiaochuan, jiaqi01, jinliang, liulu06, wangjingjing03,
xucong, guozhenhua, zhaoyaqian}@ieisystem.com, gongxiaoli@nankai.edu.cn,
lirengangx@gmail.com, fanbaoyu@ieisystem.com

A. More details about *DropletVideo-10M*.

A large proportion of videos in existing video generation datasets, such as OpenVid-1M [21], Open-Sora-Plan [18], and Panda-70M [6], primarily focus on object movements within frames while lacking camera motions. We first filtered approximately 600K high-quality spatio-temporal video clips from OpenVid-1M [21], MiraData[16], and Pexels [22]. However, this amount of data is insufficient to train a foundation video generation model. Consequently, we construct a dataset from scratch which incorporates both object movement and dynamic camera viewpoint changes. Furthermore, existing video captions, serving as textual labels and metadata, often fail to account for spatio-temporal consistency. We address this limitation by enhancing caption quality in our dataset, ensuring a more comprehensive representation of motion dynamics.

To ensure that the videos in our dataset are both realistic and practical, we construct the dataset using existing video sources, including movies, short films, VLOGs, and similar content. However, these videos are typically complex, often comprising multiple scenes, which makes them impractical for current video generation tasks. To address this, we segment the videos and selectively retain those that properly for video generation training. Specially, we focus on the videos feature both object motion and camera movement. To accomplish this task, we propose a dataset curation pipeline, as illustrated in **Fig. 2 in the manuscript**. This pipeline consists of four key stages: video collection, video segmentation, spatio-temporal variation filtering, and the generation of spatio-temporal consistent captions.

Details regarding **Video Clip Filtering** and **Video Captioning** will be elaborated in subsequent sections.



(a) Aesthetics distribution

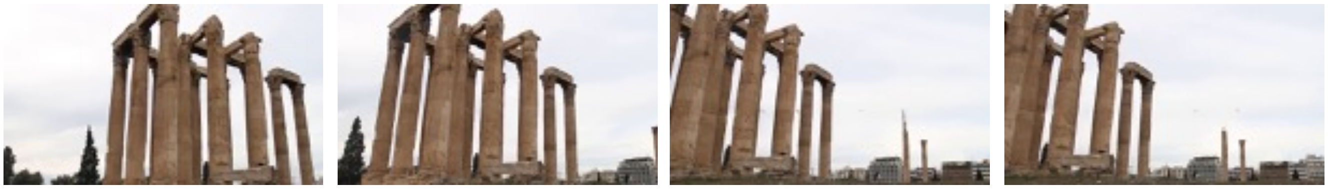
(b) Image quality distribution

Figure 1. **The aesthetics distribution and the image quality distribution of *DropletVideo-10M*.** These distributions demonstrate that our dataset achieves **high scores** in both aesthetics and image quality, indicating an overall **high-quality standard** for the dataset.

A.1. More details about Video Clip Filtering

To facilitate video generation model training, we need to select high-quality spatio-temporal video clips from the automatically segmented videos. Therefore, we developed a novel classification model, which classifies camera motion types based on the observed motion magnitude and style. We define four primary categories: (C1) camera orbiting or target self-rotation, (C2) local horizontal or vertical tilting, (C3) camera tracking a moving target, and (C4) linear camera motion. Additionally, we classify clips with static or near-static camera movement as (C5) and those edited using

[*] Equally contributing authors.
[†] Corresponding authors.



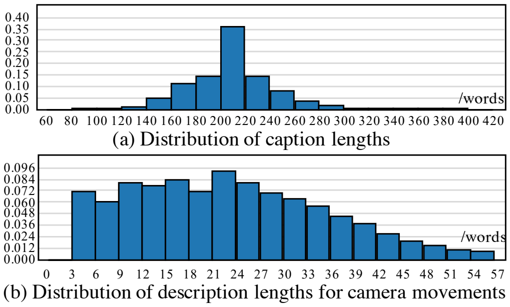
ShareGPT4Video-8B: The video portrays a static scene of ancient stone pillars, likely from historical monuments or temples, juxtaposed against a modern urban setting. The pillars, eroded by time, are linearly arranged and serve as the focal point amidst a blend of contemporary and industrial architecture. An overcast sky casts a soft glow, highlighting the contrast between ancient and modern elements. The video maintains a stationary perspective with no environmental or positional changes, nor any alterations in the scene's objects. The consistent lighting and composition suggest a snapshot of a moment, devoid of visual progression or variation.

MA-LMM-vicuna7B: The provided image is a collection of keyframes from a video, arranged in a grid view. The video captures the ancient ruins of the Acropolis of Athens, showcasing the various columns and supports that make up the structure of the Acropolis. The video also features the Parthenon, one of the most iconic landmarks of the Acropolis of Athens. The Acropolis is an ancient citadel located on a high rocky hill above the city of Athens, Greece.

ShareCaptioner-video: The video depicts a tranquil and undisturbed scene of classical ruins, featuring a row of tall and ancient pillars that have withstood the test of time. These pillars vary in height and show signs of disrepair and weathering, with ornate capitals on top, indicating they were part of a grand structure. The background of the scene is a gloomy sky, with a hint of vegetation on the right, suggesting that the ruins may be located near a park or a natural environment. As the video progresses, the camera slightly moves to the right, causing a subtle change in perspective, and the view of the ruins varies accordingly. Despite these slight variations, the pillars, sky, and vegetation remain constant, maintaining the serene atmosphere of the scene. Throughout the video, there are no obvious movements, actions, or changes, preserving the tranquility and calm around the ancient columns.

InternVL2-8B: The video showcases ancient ruins, particularly a series of towering columns that are part of a historical building, possibly a temple. The camera starts with close-up shots of the ruins, focusing on the columns, which appear in shades of blue against a backdrop of faint mist. As the camera moves, it offers a panoramic view of the ruins, highlighting the impressive architecture and the clarity of the columns. The footage captures various angles, from frontal to side views, revealing how the towering columns stand in stark contrast to the shorter ruins and the landscape. The camera continues to pan right, with modern buildings in the distance gradually coming into view. The video comes to a halt when the lens focuses on the column closest to the modern buildings. The entire video creates a sense of depth and space, with the cityscape beyond the archaeological site faintly visible in the background, suggesting the juxtaposition of history and modernity. Each shot captures more of the surrounding scenery as it pans to the right, revealing more columns and remnants of the ancient ruins.

Figure 2. **Captions generated by the fine-tuned models,** including InternVL2-8B[7, 8], ShareGPT4Video-8B[5], ShareCaptioner-video[5], and MA-LMM[12]. InternVL2-8B[7, 8] captures intricate camera work and narrative elements with high efficacy.



(a) Distribution of caption lengths

(b) Distribution of description lengths for camera movements

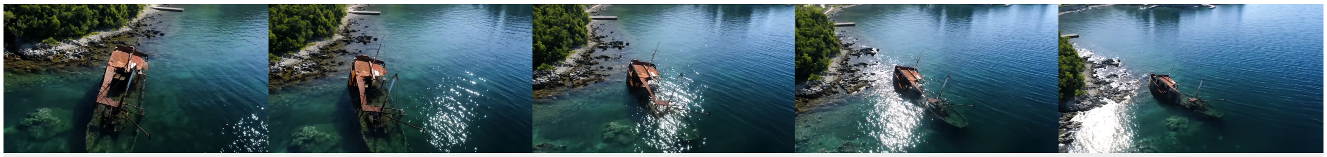
Figure 3. Word counts of captions in *DropletVideo-10M*.

software, such as transitions or artificial effects, as (C6). To ensure high-quality data, we exclude most C5 clips and all C6 clips from *DropletVideo-10M*.

To automate this process, we manually label 20,000 video clips and train a classification model based on the Video Swin Transformer[20]. We use this model to identify and categorize clips belonging to the four primary motion types, forming a spatio-temporal-aware video dataset. Additionally, we included a small proportion (less than 5%) of aesthetically pleasing and high-quality videos from class C5, as these clips contribute to enhancing the overall quality of video generation.

Next, we refine the dataset by selecting high-quality videos based on aesthetics and image quality. We utilize the publicly available LAION aesthetics model [26] to compute aesthetic scores and the DOVER-Technical model [31] to evaluate image quality. Only clips surpassing predefined thresholds are retained. The distributions of aesthetics and image quality scores for *DropletVideo-10M* are illustrated in Fig. 1. Notably, nearly 95% of clips achieve an aesthetic score above 3.5, while approximately 78% exceed a score of 4.0 in image quality, underscoring the dataset's high visual fidelity.

Additionally, we have added supplementary statistics to further illustrate it's strengths in camera motion. Fig. 3(a) depicts the 045 distribution of caption lengths for videos,



Caption: This video showcases an abandoned ship moored in a tranquil sea, surrounded by lush green vegetation and a rocky coastline. The footage is captured from a high-altitude vantage point, revealing the detailed structure of the ship and its surroundings.
As the video begins, the camera zooms in on an old, rusted ship with a brown hull, tilting towards the shore. The ship is equipped with two long masts, devoid of sails. The ship is encircled by clear seawater, which reflects the sunlight in a sparkling array, displaying varying shades of blue and green.
With the movement of the camera, the rocky coastline to the left of the ship comes into view, lined with green vegetation and scattered with small stones. The shoreline extends into the distance, meeting the sea.
Throughout the video, the ship remains stationary as the camera gradually pulls back to reveal the broader environment. To the right of the ship lies an open expanse of sea, calm and serene, with the faint outlines of other ships visible in the distance.
The entire video conveys a sense of tranquility with a touch of desolation, contrasting the ship's dilapidation with the vitality of the natural surroundings.

Caption: This video depicts a fantastical forest scene, where a small figure dressed in white is seen walking through a lush green forest. At the beginning of the video, the camera focuses on the depths of the forest, revealing a small figure in white moving from the right to the left side of the screen. Surrounding him are dense green plants, including tall trees and low shrubs. In the background, sunlight filters through the leaves onto the ground, creating a serene and mysterious atmosphere.
As the video progresses, the camera slowly pans to the left, with the small figure continuing to walk forward, revealing more details of the trees and vegetation in the background. There are also some large mushrooms with vibrant orange and red colors, adding a splash of brightness to the scene.
In the latter half of the video, the camera continues to move left, with the figure gradually exiting the frame, while the forest landscape in the background becomes even clearer. It is evident that there are rocks and moss-covered boulders, which add to the natural beauty of the forest.
The entire video, through its slow camera movement, evokes a sense of exploring an unknown world, offering a tranquil and mysterious journey through the forest.

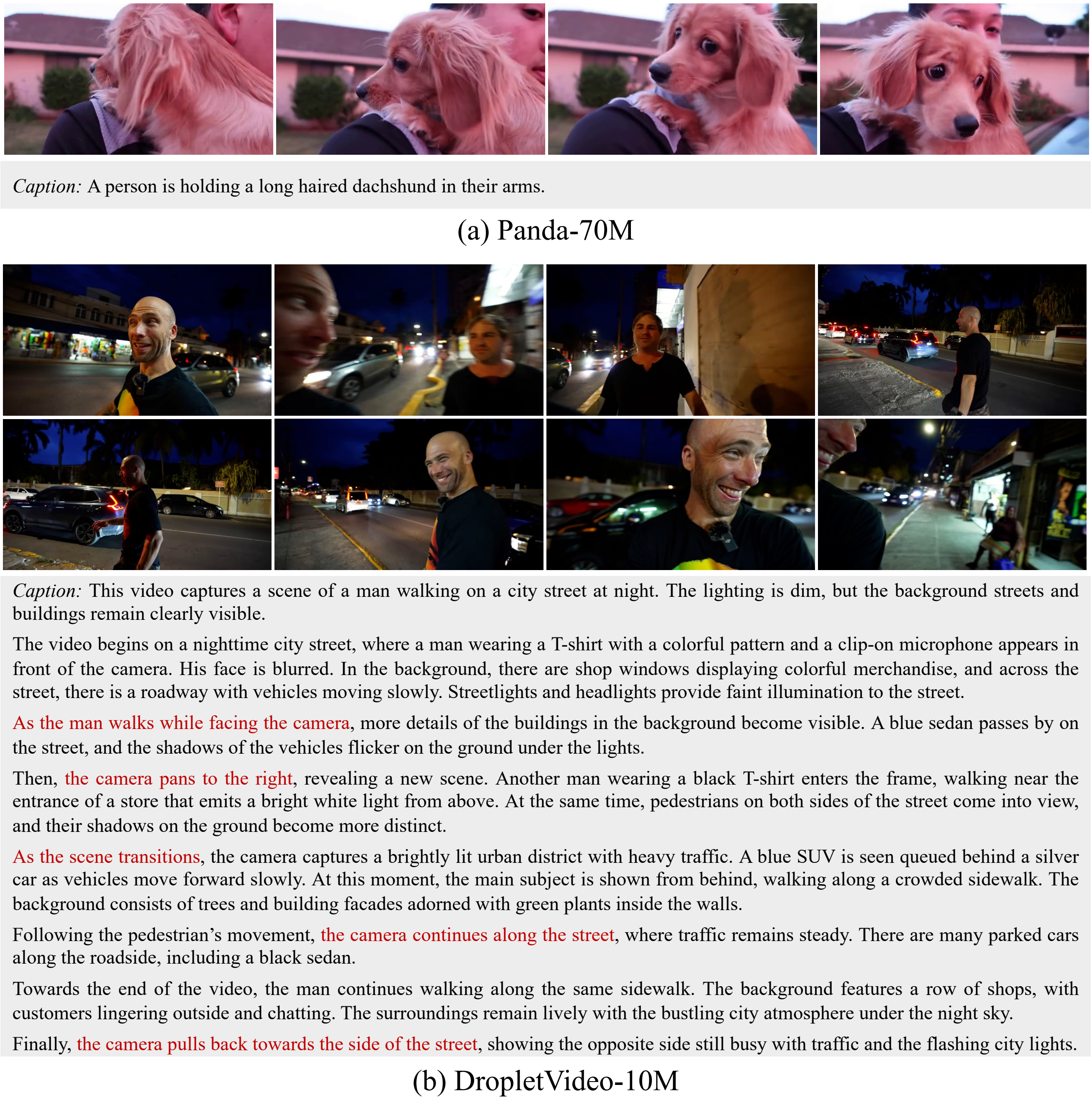
Figure 4. **Results of the fine-tuned video captioning model.** In the prompts, descriptions related to camera motions are highlighted in red. It is evident from the training samples that the camera undergoes multiple motion changes. Moreover, the scene details in the videos are clearly described and accurately followed as the camera moves. These high-density informational text captions significantly enhance the spatio-temporal semantics of the videos. Consequently, our video captions in the *DropletVideo-10M* dataset provide enriched guidance for training video generation models.

while (b) specifically shows the distribution of words related to camera motion descriptions. On average, *DropletVideo-10M* utilizes approximately 24.2 words to describe camera motion changes in relevant samples, surpassing the lengths of the entire caption in some datasets. This underscores the substantial camera motion content within *DropletVideo-10M*, providing a strong foundation for model training in this domain. Future work will involve designing models to more precisely quantify the richness of plot and camera motion.

A.2. More details about Video Captioning

We employ a video-to-text model to generate captions for video clips, reducing the need for extensive human labor. However, existing video-to-text models typically produce brief descriptions, which are insufficient for ensuring spatio-temporal consistency in our video generation task. To address this, we first curate a dataset of videos with captions that provide comprehensive descriptions of objects, scenes, and visual transitions, with a particular emphasis on camera movements and their effects. Subsequently, we utilize GPT-4 to correct grammatical and spelling errors in the captions, thereby creating a dataset suitable for instructionally supervised fine-tuning of multimodal models. Based on this dataset, we fine-tune several open-source multimodal models, each containing approximately 7–9 billion parameters, including InternVL2-8B[7, 8], ShareGPT4Video-8B[5], ShareCaptioner-Video[5], and MA-LMM[12]. Fig. 2 illustrates the outcomes of these models.

We evaluate these models using human expert ratings. Both InternVL2-8B and ShareCaptioner demonstrate strong



Caption: A person is holding a long haired dachshund in their arms.

(a) Panda-70M

Caption: This video captures a scene of a man walking on a city street at night. The lighting is dim, but the background streets and buildings remain clearly visible.

The video begins on a nighttime city street, where a man wearing a T-shirt with a colorful pattern and a clip-on microphone appears in front of the camera. His face is blurred. In the background, there are shop windows displaying colorful merchandise, and across the street, there is a roadway with vehicles moving slowly. Streetlights and headlights provide faint illumination to the street.

As the man walks while facing the camera, more details of the buildings in the background become visible. A blue sedan passes by on the street, and the shadows of the vehicles flicker on the ground under the lights.

Then, the camera pans to the right, revealing a new scene. Another man wearing a black T-shirt enters the frame, walking near the entrance of a store that emits a bright white light from above. At the same time, pedestrians on both sides of the street come into view, and their shadows on the ground become more distinct.

As the scene transitions, the camera captures a brightly lit urban district with heavy traffic. A blue SUV is seen queued behind a silver car as vehicles move forward slowly. At this moment, the main subject is shown from behind, walking along a crowded sidewalk. The background consists of trees and building facades adorned with green plants inside the walls.

Following the pedestrian's movement, the camera continues along the street, where traffic remains steady. There are many parked cars along the roadside, including a black sedan.

Towards the end of the video, the man continues walking along the same sidewalk. The background features a row of shops, with customers lingering outside and chatting. The surroundings remain lively with the bustling city atmosphere under the night sky.

Finally, the camera pulls back towards the side of the street, showing the opposite side still busy with traffic and the flashing city lights.

(b) DropletVideo-10M

Figure 5. **The *DropletVideo-10M* dataset features diverse camera movements, long-captioned contextual descriptions, and strong spatio-temporal consistency.** (a) Existing datasets, such as Panda-70M [6], place less emphasis on camera movement and contain relatively brief captions. (b) In contrast, *DropletVideo-10M* consists of spatio-temporal videos that incorporate both camera movement and event progression. Each video is paired with a caption that conveys detailed spatio-temporal information aligned with the video content, with an average caption length of 206 words. The spatio-temporal information is highlighted in red in the figure.

performance, excelling at generating detailed and precise descriptions of camera movements while maintaining semantic richness and coherence. However, ShareCaptioner-Video exhibits significantly reduced efficiency due to its sliding captioning and clip summarization strategy, which requires distinct descriptions for each sampled frame, leading to more frequent LLM invocations. Balancing efficiency and performance, we selected the fine-tuned InternVL2-8B for large-scale caption generation in the *DropletVideo-10M* dataset. To further enhance captioning quality, we refined the video captioning model based on InternVL2-8B using LoRA [14]. This improved model generates highly detailed descriptions that accurately capture interactions caused by lens changes, including camera movements, various transitions, and content shifts, thereby providing precise training data for video generation models.

Fig. 4 presents two complete samples, illustrating that the generated captions comprehensively describe camera operations and the visual transitions induced by motion. Additionally, we ensured that descriptions include sufficient details about lighting, style, and atmosphere of objects and backgrounds, thereby offering richer guidance for model training. Each video segment is annotated with captions averaging 206 words in length, ensuring a high level of detail and descriptive accuracy.

A.3. Demonstration of DropletVideo-10M

As depicted in Fig. 5, *DropletVideo-10M* features a diverse array of camera viewpoint transformations that preserve plot continuity, adhering to the integral spatio-temporal consistency criteria in video content. Additionally, the associated prompts offer comprehensive information, with a focus on detailing camera movements and their effects on objects and scenes. From this perspective, *DropletVideo-10M* supplies more granular supervisory signals, facilitating the model's acquisition of visual and semantic knowledge, as well as proficiency in camera operation during video generation.

B. More details about *DropletVideo*.

Open-source video generation models generally exhibit poorer performance compared to closed-source models, particularly when both camera angles and scene movements occur. These models face challenges in preserving temporal and spatial consistency, factoring in camera angles and object movement, and lack mechanisms to regulate plot or camera speed. Consequently, we introduce the *DropletVideo* model, the architecture of which is detailed in **Fig. 3 in the manuscript**.

B.1. Preliminary Overview of Diffusion Models

The proposed *DropletVideo* is developed and trained utilizing diffusion model (DM)[9]. The essence of a DM involves generating samples from a distribution by reversing a gradual noising process. This process initiates with a noisy input, x_T, which is usually Gaussian noise, and sequentially produces less noisy samples, $x_{T-1}, x_{T-2}, \ldots$, culminating in the final sample, x_0. The timestep t is used to indicate the noise level. x_t represents a combination of the original signal x_0 and added noise ϵ.

During the diffusion phase, the model progressively adds noise to the data, increasing in intensity until the original data is fully transformed into Gaussian noise. Given a real data distribution $x_0 \sim q(x)$, and it is sampled T times to add Gaussian noise. The variation schedule of the noise is defined as a_t, and the data thus sampled is denoted as x_t, where $t \in [1, T]$. The process obeys a Markov chain, and after a reparameterization trick, the model can directly obtain any intermediate state, and the sampling formula for x_t is $q(x_t) = N(x_t; \sqrt{\bar{a}_t} x_0, (1 - \sqrt{\bar{a}_t}) I)$, where $\bar{a}_t = \prod_{i=1}^{t} a_i$.

Conversely, during the denoising phase, the model learns the real data distribution from the standard Gaussian noise $p(x_T)$, where $p(x_T) = \mathcal{N}(x_T; 0, I)$. The DM is trained to generate a successively denoised x_{t-1} from x_t. Ho et al. [13] define the model as a function $\epsilon_\theta(x_t, t)$ that estimates the noise component in the noisy sample x_t. The noise prediction function $\epsilon_\theta(x_t, t)$ is usually obtained by designing a U-Net network stacked with residual networks. The optimization objective is then defined as $\|\epsilon_\theta(x_t, t) - \epsilon_t\|^2$, where ϵ_t represents the sampled noise at time t and serves as the ground truth.

To mitigate the high computational and resource demands of conventional diffusion models in generating high-dimensional data, a series of latent diffusion models (LDMs) [24] has been introduced. A LDM employs a pre-trained perceptual compression model consisting of an encoder ε and a decoder D [4, 23]. This integration allows the diffusion process to transfer from the high-dimensional pixel space to the low-dimensional latent space, thereby enabling learning in the latent representation domain. The objective function of the LDM is $L_{LDM} = E_{\varepsilon(x_0), t, \epsilon_\theta \sim N(0,I)} [\|\epsilon_t - \epsilon_\theta(z_t, t)\|^2]$, where z_t is the output of the encoder.

Drawing inspiration from 3D Variational Autoencoders [32], *DropletVideo* model encodes video frames into the latent space using three-dimensional convolutions, capturing both spatial and temporal dimensions. Additionally, we incorporate the Multi-Modal Diffusion Transformer (MMDiT) model [10]. This integration permits the model to function autonomously within the representation spaces of text and video, while also accounting for their inter-dependencies, thereby facilitating enhanced information transfer and synthesis.

C. More demonstrations and discussions in the experiment.

C.1. Qualitative Evaluation

C.1.1. Integral Spatio-temporal Consistency

Dynamic Scene Generation with Integral Spatio-temporal Consistency. *DropletVideo* focuses on integral spatio-temporal consistency during video generation. It addresses the spatial distortion issues caused by camera movement, ensuring smooth plot progression during camera movement and the spatio-temporal consistency of objects within the scene. More importantly, in the development of a video scenario, the emerging scenes do not affect the behavior of the original video objects. Fig. 6 exemplifies the integral spatio-temporal consistency. It is evident that *DropletVideo* can maintain the continuity of the original plot while new plots enter the video.

High controllability of Emerging objects. To further validate *DropletVideo*'s capability in generating videos with integral spatio-temporal consistency, we conducted ablation studies focusing on the driving prompts. By modifying only the final sentences of the prompts while keeping the rest unchanged, we assessed the system's precision in controlling the characteristics of emerging objects, as shown in Fig. 7. The resulting videos clearly demonstrate *DropletVideo*'s exceptional ability to accurately translate textual descriptions into visual elements, ensuring a high degree of fidelity to the specified attributes. This highlights *DropletVideo*'s remarkable control over the emergence and detailed features of objects within the generated videos.

C.1.2. 3D Consistency

Trained on the large-scale spatio-temporal dataset, *DropletVideo-10M*, *DropletVideo* exhibits remarkable 3D consistency, as illustrated in Fig. 8. In the top example, the camera rotates around a snowflake, maintaining stringent consistency for both the background and the snowflake from various angles, while preserving the snowflake's intricate details across multiple perspectives. In the bottom example, the camera performs an arc shot, projecting the same object. Despite not being specifically designed for arc shots, *DropletVideo* effectively maintains the insect's 3D consistency over a broad range of rotation angles, demonstrating robust spatial 3D continuity.

C.1.3. Controllable Motion Intensity

DropletVideo manipulates the rate of plot progression and camera angle transitions through the adjustment of a motion control parameter. In the given example, enhancing this parameter allows a video of identical duration to accommodate more plot elements. Fig. 9 displays the video generation results under various motion control parameters using the same text-image input. Under the setting of $M = 8$, the camera's movement is noticeably more pronounced than at $M = 12$ and $M = 16$, where the snowflake is presented with a broader range of perspectives. The motion density decreases as the M escalates from 8 to 16, confirming that a lower M results in a video with more drastic camera variations. This evidence suggests that *DropletVideo* can adeptly regulate the playback speed of the content while maintaining semantic accuracy.

C.1.4. Camera Motion

DropletVideo demonstrates versatile camera motion generation capabilities including various fundamental movement types, as visualized in Fig. 10. The system produces cinema-standard motions including right/left trucking, vertical pedestal movement, tilt adjustment, axial dollying, and composite pan-tilt operations.

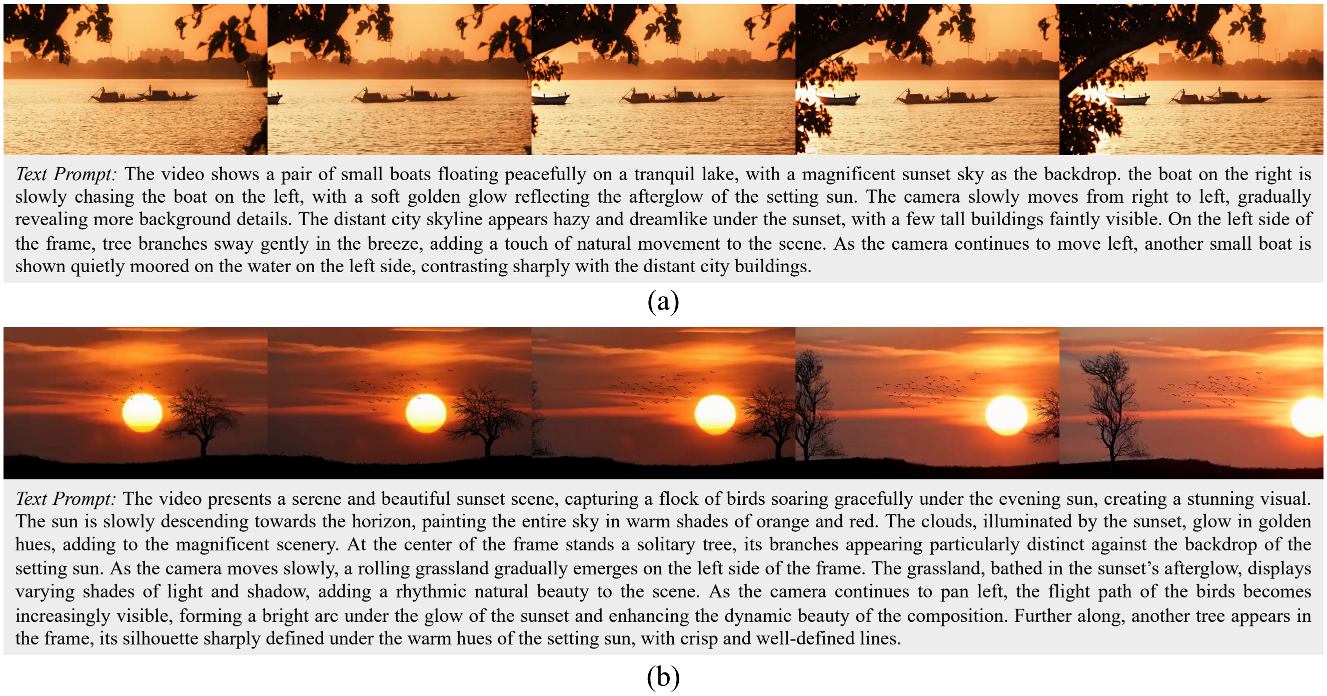
Camera truck right. Fig.10 (a) illustrates precise truck right control through foliage dynamics. Beginning with a micro view of dual leaves, the system executes text-guided trucking movement where left leaf edges fade proportionally as right venation textures emerge. Focus transitions between flowing and static droplets maintain optical continuity, with refractive stability persisting despite background bokeh deformation that adheres to lens physics.

Camera truck left. The riverbank case in Fig. 10 (b) showcases environmental expansion during leftward movement. Initial partial riverbank frames progressively incorporate complete stone formations and canopy structures, maintaining geometric coherence between existing and generated elements. Vehicle mud stains preserve spatial consistency while water ripples develop accurate motion parallax relative to camera speed. Dynamic light refraction on aquatic surfaces replicates real fluid behavior, particularly in water droplet translucency during splash events.

Camera Pedestal down. Vertical control in botanical close-ups in Fig. 10 (c) manifests through synchronized plant revelation. Descending motion coordinates with stem texture emergence, where curled leaves gradually unfurl following botanical growth patterns. Background vegetation blur intensifies proportionally to focal plane descent, matching professional lens depth-of-field characteristics. Waxy surface highlights migrate smoothly across the leaves, preserving material authenticity during viewpoint transitions.

Camera Tilt up. The architectural validation in Fig.10 (d) confirms 3D spatial awareness during upward tilting. The spiral staircase geometry remains intact with stable railing spacing and curvature radii, while newly revealed decorative elements scale according to perspective principles. Color constancy persists across lighting variations, evidenced by consistent carpet saturation and wall temperature. Chandelier glow attenuation follows inverse-square law principles, with wall decorations maintaining physically accurate diffuse reflections.

Camera Dolly in. The snowscape progression in Fig. 10



Text Prompt: The video shows a pair of small boats floating peacefully on a tranquil lake, with a magnificent sunset sky as the backdrop. the boat on the right is slowly chasing the boat on the left, with a soft golden glow reflecting the afterglow of the setting sun. The camera slowly moves from right to left, gradually revealing more background details. The distant city skyline appears hazy and dreamlike under the sunset, with a few tall buildings faintly visible. On the left side of the frame, tree branches sway gently in the breeze, adding a touch of natural movement to the scene. As the camera continues to move left, another small boat is shown quietly moored on the water on the left side, contrasting sharply with the distant city buildings.

(a)

Text Prompt: The video presents a serene and beautiful sunset scene, capturing a flock of birds soaring gracefully under the evening sun, creating a stunning visual. The sun is slowly descending towards the horizon, painting the entire sky in warm shades of orange and red. The clouds, illuminated by the sunset, glow in golden hues, adding to the magnificent scenery. At the center of the frame stands a solitary tree, its branches appearing particularly distinct against the backdrop of the setting sun. As the camera moves slowly, a rolling grassland gradually emerges on the left side of the frame. The grassland, bathed in the sunset's afterglow, displays varying shades of light and shadow, adding a rhythmic natural beauty to the scene. As the camera continues to pan left, the flight path of the birds becomes increasingly visible, forming a bright arc under the glow of the sunset and enhancing the dynamic beauty of the composition. Further along, another tree appears in the frame, its silhouette sharply defined under the warm hues of the setting sun, with crisp and well-defined lines.

(b)

Figure 6. ***DropletVideo* facilitates the generation of videos that maintain integral spatio-temporal consistency.** New objects or scenes introduced via camera movement are seamlessly integrated and interact logically with the pre-existing scenes. In video (a), as the camera moves, a new boat appears on the lake, the boat on the right of the original two boats continues to slowly chase the boat on the left, and the leaves on the shore still sway gently in the breeze. In video (b), as the camera moves left, the tree called for in the text prompt successfully appears in the shot, the original flock of birds continues to fly, and the grass and sky show continuity as the camera moves.

(e) demonstrates axial movement precision. Forward camera motion proportionally reveals architectural details: initially obscured red doors gradually restore surface textures under natural light decay, while window reflections adjust intensity with viewing distance. Pine trees maintain spatial reference integrity, their parallax displacements creating authentic depth gradients between foreground snow paths and background vegetation.

Camera Pan right And Tilt up. Composite motion control in Fig. 10 (f) achieves seamless transition from lakeside panning to skyward tilting. Initial rightward movement preserves accurate spatial relationships between water glare and mountain reflections, with snow distribution transitioning naturally. During axis transition, lake area proportion decreases geometrically while emerging cloud formations maintain pattern continuity. Altitude-dependent lighting differentiation enhances realism, where high-altitude cloud translucency contrasts distinctly with low-altitude texture density.

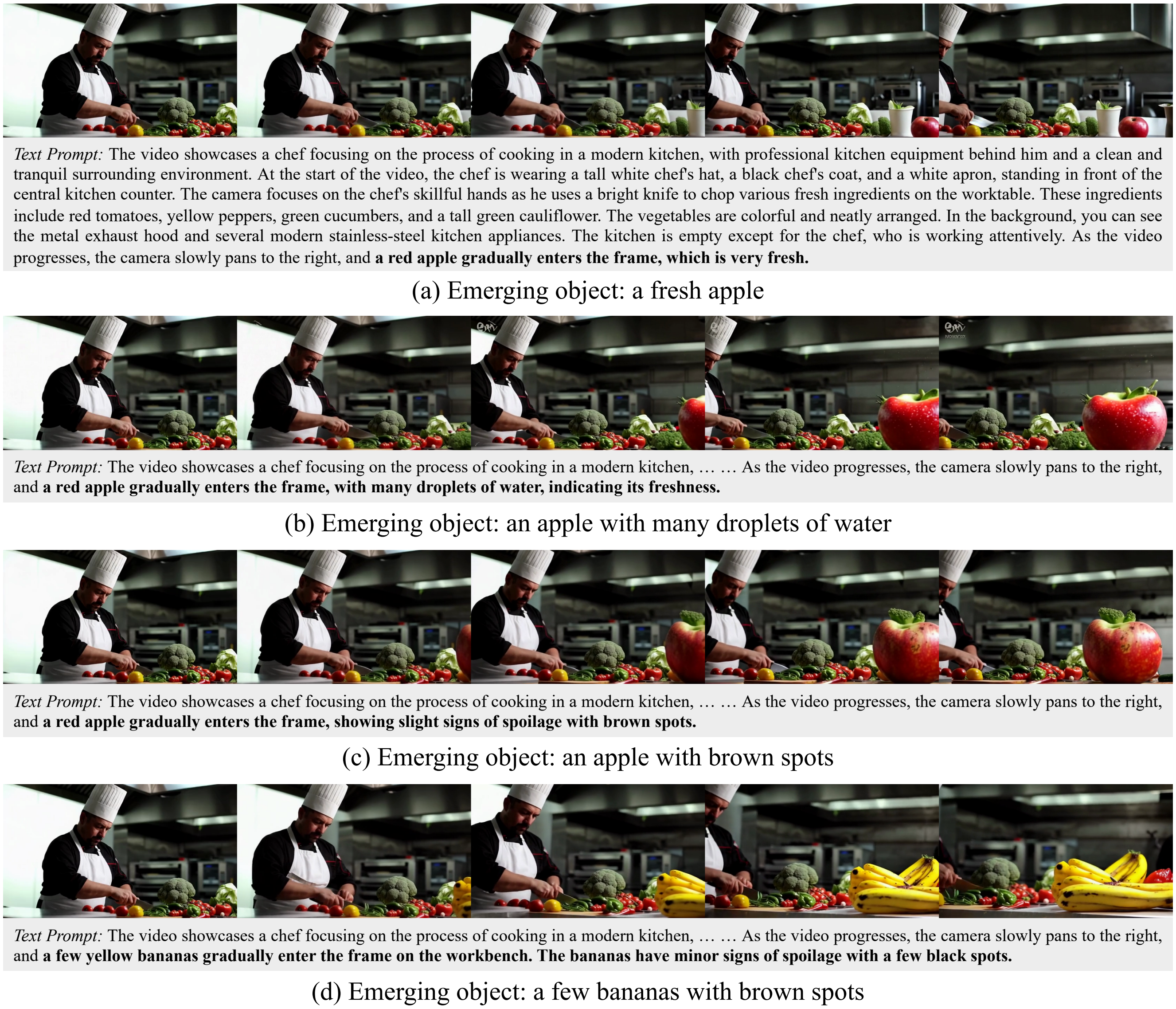
C.1.5. Comparison of our *DropletVideo* with existing Models

To better demonstrate the cumulative spatiotemporal consistency of *DropletVideo*, we have selected several industry-recognized video generation models for comparison, including Hailuo [11], Kling v1.6 [17], Gen-3 [25], Vidu [29], Vivago [30], Qingying [2], CogVideoX-Fun [3], and WanX [28]. Out of the compared models, only CogVideoX-Fun and WanX are open-source, similar to our approach, whereas the remaining models are closed-source.

We conducted comparisons using examples from various scenarios mentioned earlier, such as boat, kitchen, lake, snow, staircase, and sunset, as shown in Fig. 11 - 16.

The examples of boat, sunset, and kitchen are particularly effective for evaluating cumulative spatiotemporal consistency, as they involve diverse spatial transformations and detailed descriptions of target features. From these examples, we observe that WanX [28] and Kling v1.6 [17] perform relatively well. However, no single model excels across all these scenarios. In contrast, our algorithm consistently demonstrates superior spatio-temporal consistency across these examples. For instance, as depicted in Fig. 12, *DropletVideo* successfully produces a video where the camera rotation is precisely captured, simultaneously portraying the chasing interaction between the two boats. This level of detail and accuracy is beyond the capabilities of some other models, which struggle to generate such a scene with the same fidelity.

The scenarios of snow, staircase, and lake highlight



Text Prompt: The video showcases a chef focusing on the process of cooking in a modern kitchen, with professional kitchen equipment behind him and a clean and tranquil surrounding environment. At the start of the video, the chef is wearing a tall white chef's hat, a black chef's coat, and a white apron, standing in front of the central kitchen counter. The camera focuses on the chef's skillful hands as he uses a bright knife to chop various fresh ingredients on the worktable. These ingredients include red tomatoes, yellow peppers, green cucumbers, and a tall green cauliflower. The vegetables are colorful and neatly arranged. In the background, you can see the metal exhaust hood and several modern stainless-steel kitchen appliances. The kitchen is empty except for the chef, who is working attentively. As the video progresses, the camera slowly pans to the right, and **a red apple gradually enters the frame, which is very fresh.**

(a) Emerging object: a fresh apple

Text Prompt: The video showcases a chef focusing on the process of cooking in a modern kitchen, … … As the video progresses, the camera slowly pans to the right, and **a red apple gradually enters the frame, with many droplets of water, indicating its freshness.**

(b) Emerging object: an apple with many droplets of water

Text Prompt: The video showcases a chef focusing on the process of cooking in a modern kitchen, … … As the video progresses, the camera slowly pans to the right, and **a red apple gradually enters the frame, showing slight signs of spoilage with brown spots.**

(c) Emerging object: an apple with brown spots

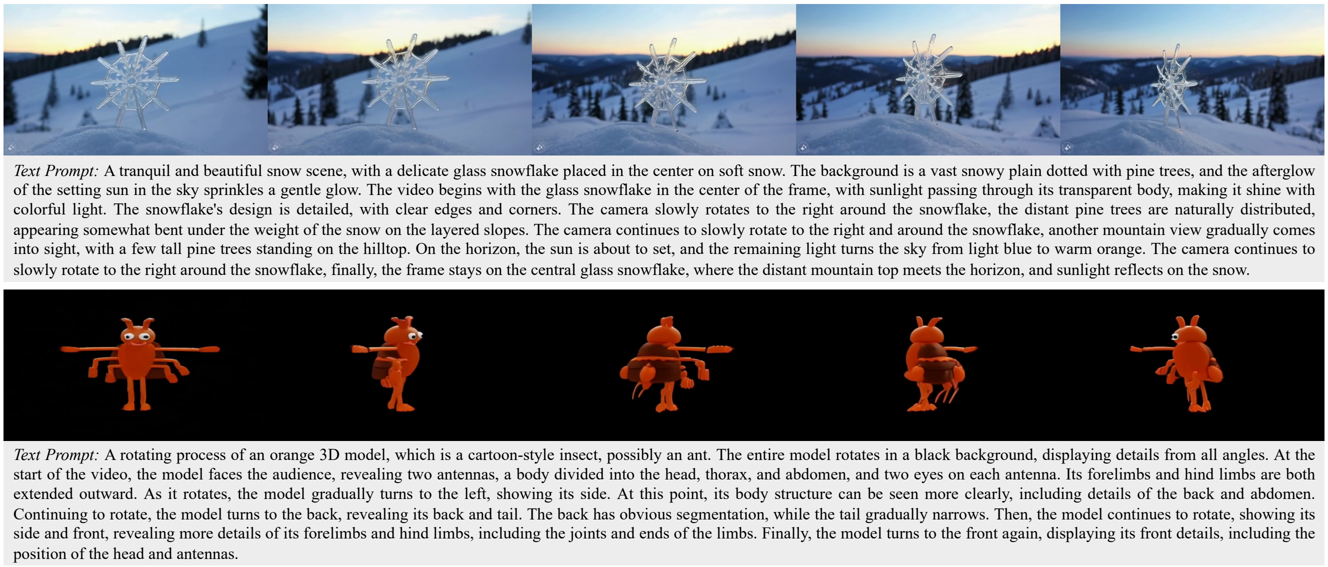
Text Prompt: The video showcases a chef focusing on the process of cooking in a modern kitchen, … … As the video progresses, the camera slowly pans to the right, and **a few yellow bananas gradually enter the frame on the workbench. The bananas have minor signs of spoilage with a few black spots.**

(d) Emerging object: a few bananas with brown spots

Figure 7. ***DropletVideo* demonstrates advanced controllability in generating scenes where new objects emerges due to camera movement.** In video (a), as the camera pans right, the red apple specified in the prompt appears seamlessly, while the chef continues cooking, illustrating smooth integration of new objects. Video (b) showcases the system's ability to handle detailed descriptions, as the prompt's depiction of an apple with water droplets is rendered accurately, highlighting complex textures. In video (c), a prompt modification adds brown spots to the apple, which are visibly integrated, showing dynamic visual adjustments. Finally, in video (d), the prompt changes the apple to bananas, and the system adeptly features bananas, demonstrating versatility and precision in object transformation.

DropletVideo's exceptional camera movement capabilities. Among the other algorithms, Kling v1.6 [17] performs better, yet others fall short. Our algorithm, however, performs exceptionally well, closely adhering to the instructions given in the prompt.

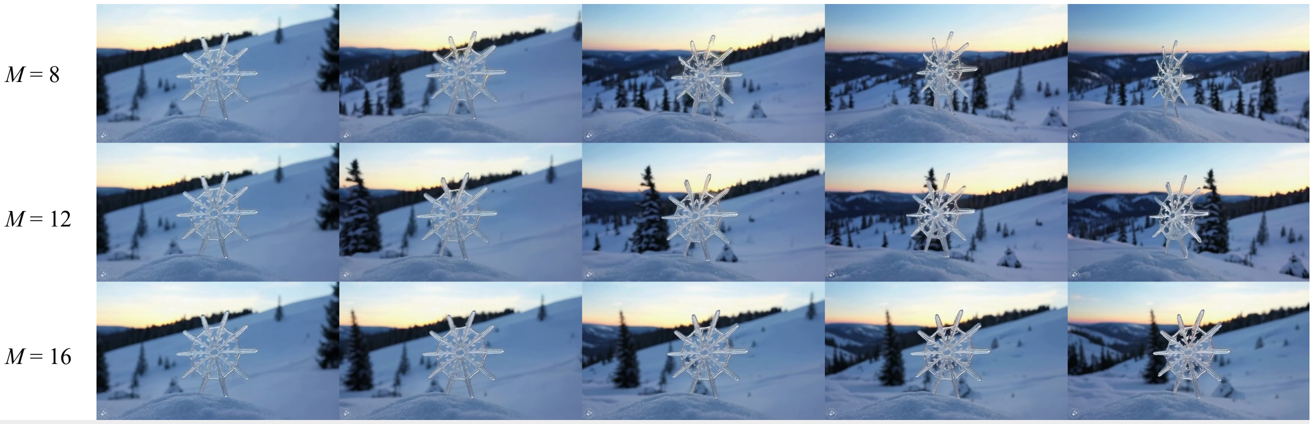
Overall, despite *DropletVideo* being an open-source model, it achieves and even surpasses the performance of existing well-known commercial generation models in terms of cumulative spatiotemporal consistency. From this perspective, we believe that *DropletVideo* holds greater promise in advancing the progress within the video generation community.



Text Prompt: A tranquil and beautiful snow scene, with a delicate glass snowflake placed in the center on soft snow. The background is a vast snowy plain dotted with pine trees, and the afterglow of the setting sun in the sky sprinkles a gentle glow. The video begins with the glass snowflake in the center of the frame, with sunlight passing through its transparent body, making it shine with colorful light. The snowflake's design is detailed, with clear edges and corners. The camera slowly rotates to the right around the snowflake, the distant pine trees are naturally distributed, appearing somewhat bent under the weight of the snow on the layered slopes. The camera continues to slowly rotate to the right and around the snowflake, another mountain view gradually comes into sight, with a few tall pine trees standing on the hilltop. On the horizon, the sun is about to set, and the remaining light turns the sky from light blue to warm orange. The camera continues to slowly rotate to the right around the snowflake, finally, the frame stays on the central glass snowflake, where the distant mountain top meets the horizon, and sunlight reflects on the snow.

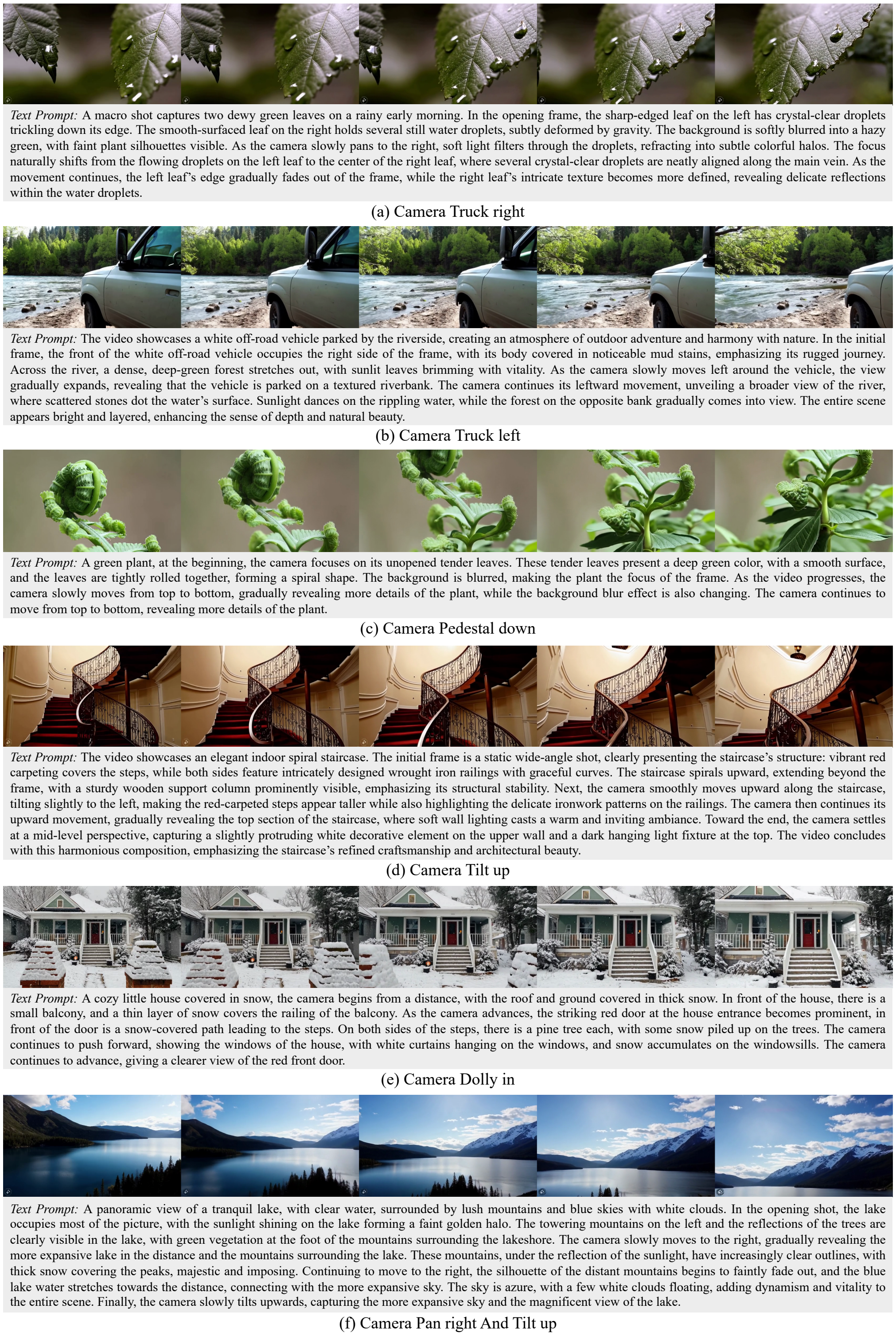
Text Prompt: A rotating process of an orange 3D model, which is a cartoon-style insect, possibly an ant. The entire model rotates in a black background, displaying details from all angles. At the start of the video, the model faces the audience, revealing two antennas, a body divided into the head, thorax, and abdomen, and two eyes on each antenna. Its forelimbs and hind limbs are both extended outward. As it rotates, the model gradually turns to the left, showing its side. At this point, its body structure can be seen more clearly, including details of the back and abdomen. Continuing to rotate, the model turns to the back, revealing its back and tail. The back has obvious segmentation, while the tail gradually narrows. Then, the model continues to rotate, showing its side and front, revealing more details of its forelimbs and hind limbs, including the joints and ends of the limbs. Finally, the model turns to the front again, displaying its front details, including the position of the head and antennas.

Figure 8. *DropletVideo* demonstrates excellent 3D consistency. In the top example, the camera moves around a snowflake, showcasing significant camera movement while maintaining the snowflake's details from multiple perspectives. In the bottom example, the camera circles around an insect, and *DropletVideo* ensures the insect's 3D consistency across a wide range of rotation angles. However, *DropletVideo* still has limitations in generating content for a full 360-degree rotation, which will be addressed in future work. Overall, these examples illustrate *DropletVideo*'s strong performance in spatial 3D consistency.



Text Prompt: A tranquil and beautiful snow scene, with a delicate glass snowflake placed in the center on soft snow. The background is a vast snowy plain dotted with pine trees, and the afterglow of the setting sun in the sky sprinkles a gentle glow. The video begins with the glass snowflake in the center of the frame, with sunlight passing through its transparent body, making it shine with colorful light. The snowflake's design is detailed, with clear edges and corners. The camera slowly rotates to the right around the snowflake, the distant pine trees are naturally distributed, appearing somewhat bent under the weight of the snow on the layered slopes. The camera continues to slowly rotate to the right and around the snowflake, another mountain view gradually comes into sight, with a few tall pine trees standing on the hilltop. On the horizon, the sun is about to set, and the remaining light turns the sky from light blue to warm orange. The camera continues to slowly rotate to the right around the snowflake, finally, the frame stays on the central glass snowflake, where the distant mountain top meets the horizon, and sunlight reflects on the snow.

Figure 9. *DropletVideo* facilitates precision control over video generation speed. Modifying the Input Speed parameter alters the movement speed of both the camera and target. In the third line, the camera motion parameter M is doubled, and the snowflake's rotation speed is substantially decreased compared to the initial setting.



Text Prompt: A macro shot captures two dewy green leaves on a rainy early morning. In the opening frame, the sharp-edged leaf on the left has crystal-clear droplets trickling down its edge. The smooth-surfaced leaf on the right holds several still water droplets, subtly deformed by gravity. The background is softly blurred into a hazy green, with faint plant silhouettes visible. As the camera slowly pans to the right, soft light filters through the droplets, refracting into subtle colorful halos. The focus naturally shifts from the flowing droplets on the left leaf to the center of the right leaf, where several crystal-clear droplets are neatly aligned along the main vein. As the movement continues, the left leaf's edge gradually fades out of the frame, while the right leaf's intricate texture becomes more defined, revealing delicate reflections within the water droplets.

(a) Camera Truck right

Text Prompt: The video showcases a white off-road vehicle parked by the riverside, creating an atmosphere of outdoor adventure and harmony with nature. In the initial frame, the front of the white off-road vehicle occupies the right side of the frame, with its body covered in noticeable mud stains, emphasizing its rugged journey. Across the river, a dense, deep-green forest stretches out, with sunlit leaves brimming with vitality. As the camera slowly moves left around the vehicle, the view gradually expands, revealing that the vehicle is parked on a textured riverbank. The camera continues its leftward movement, unveiling a broader view of the river, where scattered stones dot the water's surface. Sunlight dances on the rippling water, while the forest on the opposite bank gradually comes into view. The entire scene appears bright and layered, enhancing the sense of depth and natural beauty.

(b) Camera Truck left

Text Prompt: A green plant, at the beginning, the camera focuses on its unopened tender leaves. These tender leaves present a deep green color, with a smooth surface, and the leaves are tightly rolled together, forming a spiral shape. The background is blurred, making the plant the focus of the frame. As the video progresses, the camera slowly moves from top to bottom, gradually revealing more details of the plant, while the background blur effect is also changing. The camera continues to move from top to bottom, revealing more details of the plant.

(c) Camera Pedestal down

Text Prompt: The video showcases an elegant indoor spiral staircase. The initial frame is a static wide-angle shot, clearly presenting the staircase's structure: vibrant red carpeting covers the steps, while both sides feature intricately designed wrought iron railings with graceful curves. The staircase spirals upward, extending beyond the frame, with a sturdy wooden support column prominently visible, emphasizing its structural stability. Next, the camera smoothly moves upward along the staircase, tilting slightly to the left, making the red-carpeted steps appear taller while also highlighting the delicate ironwork patterns on the railings. The camera then continues its upward movement, gradually revealing the top section of the staircase, where soft wall lighting casts a warm and inviting ambiance. Toward the end, the camera settles at a mid-level perspective, capturing a slightly protruding white decorative element on the upper wall and a dark hanging light fixture at the top. The video concludes with this harmonious composition, emphasizing the staircase's refined craftsmanship and architectural beauty.

(d) Camera Tilt up

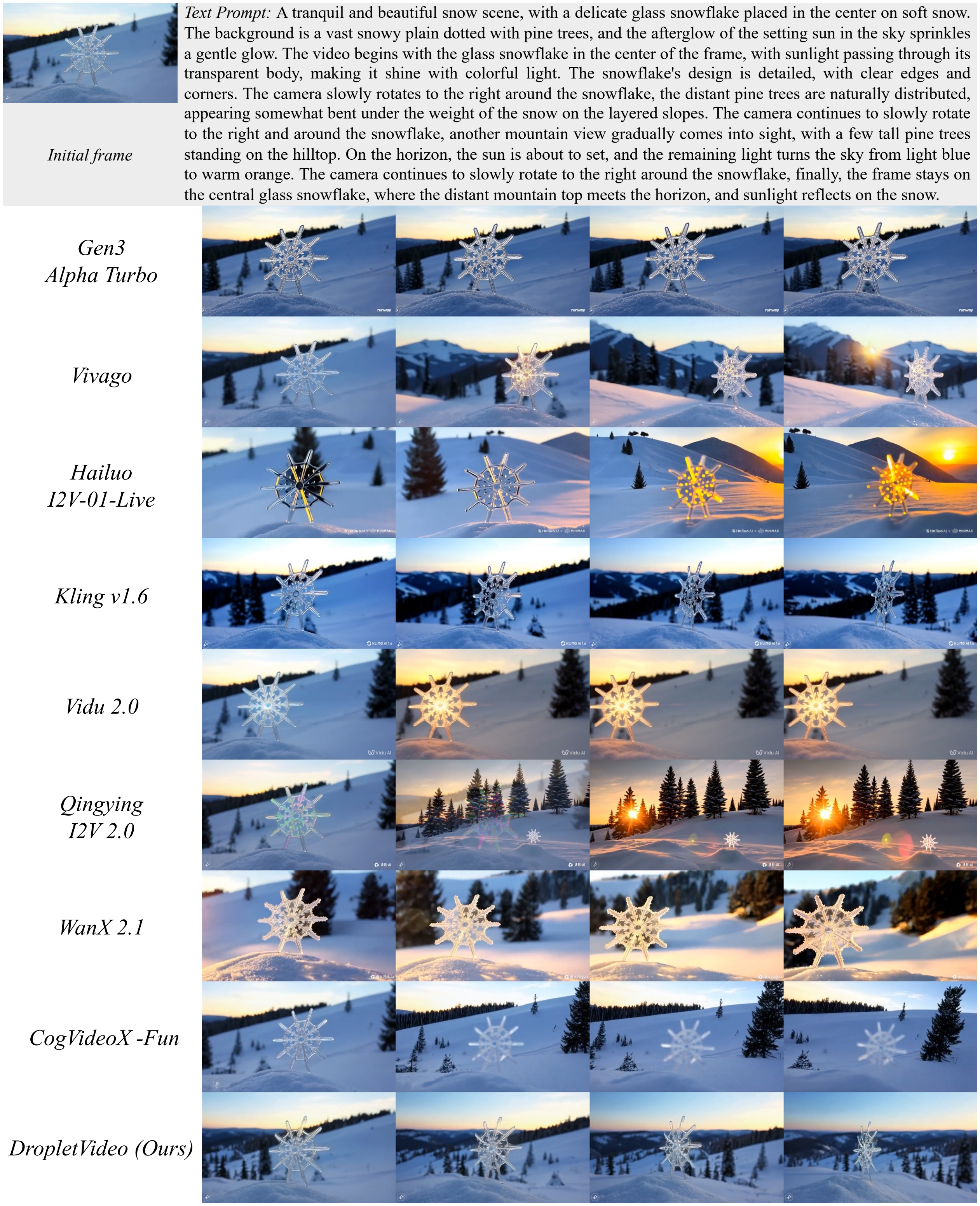
Text Prompt: A cozy little house covered in snow, the camera begins from a distance, with the roof and ground covered in thick snow. In front of the house, there is a small balcony, and a thin layer of snow covers the railing of the balcony. As the camera advances, the striking red door at the house entrance becomes prominent, in front of the door is a snow-covered path leading to the steps. On both sides of the steps, there is a pine tree each, with some snow piled up on the trees. The camera continues to push forward, showing the windows of the house, with white curtains hanging on the windows, and snow accumulates on the windowsills. The camera continues to advance, giving a clearer view of the red front door.

(e) Camera Dolly in

Text Prompt: A panoramic view of a tranquil lake, with clear water, surrounded by lush mountains and blue skies with white clouds. In the opening shot, the lake occupies most of the picture, with the sunlight shining on the lake forming a faint golden halo. The towering mountains on the left and the reflections of the trees are clearly visible in the lake, with green vegetation at the foot of the mountains surrounding the lakeshore. The camera slowly moves to the right, gradually revealing the more expansive lake in the distance and the mountains surrounding the lake. These mountains, under the reflection of the sunlight, have increasingly clear outlines, with thick snow covering the peaks, majestic and imposing. Continuing to move to the right, the silhouette of the distant mountains begins to faintly fade out, and the blue lake water stretches towards the distance, connecting with the more expansive sky. The sky is azure, with a few white clouds floating, adding dynamism and vitality to the entire scene. Finally, the camera slowly tilts upwards, capturing the more expansive sky and the magnificent view of the lake.

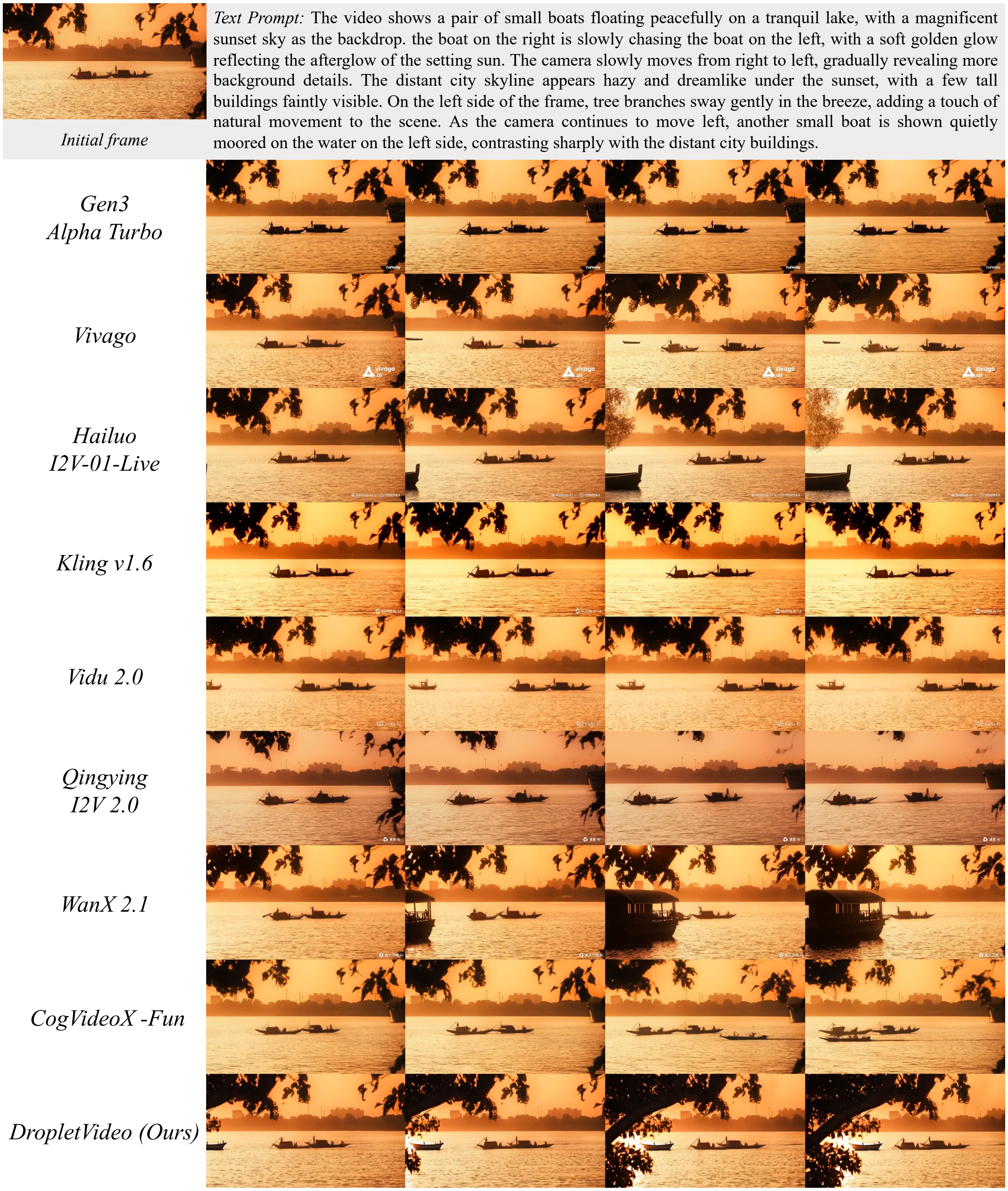
(f) Camera Pan right And Tilt up

Figure 10. ***DropletVideo* showcases its robust capabilities in generating videos with diverse camera movements.** Panels (a)-(e) illustrate the outcomes of specific camera motions: Camera Truck Right, Camera Truck Left, Camera Pedestal Down, Camera Tilt Up, and Camera Dolly In. Panel (f) presents a composite camera shot that combines Camera Pan Right and Tilt Up.



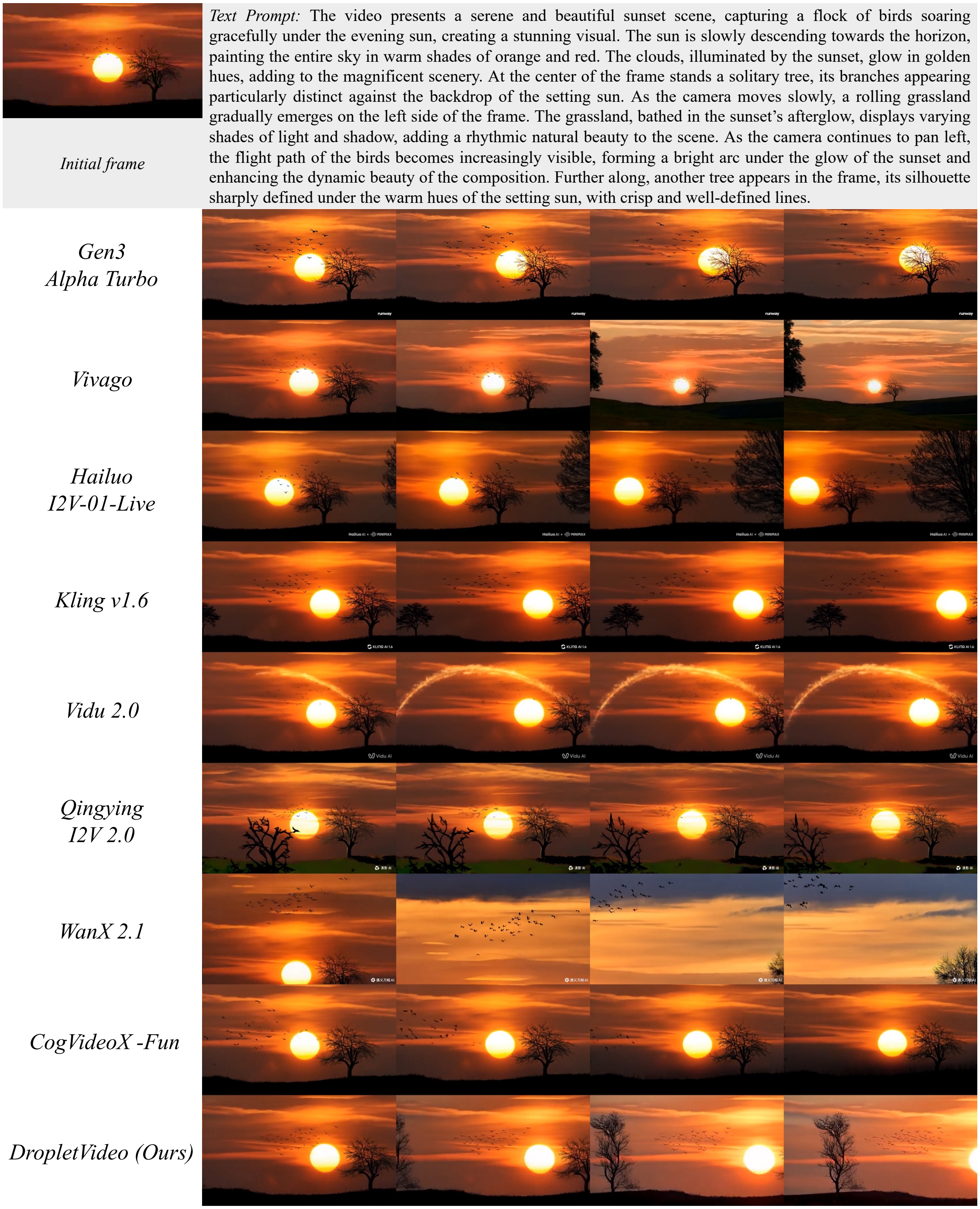
Text Prompt: A tranquil and beautiful snow scene, with a delicate glass snowflake placed in the center on soft snow. The background is a vast snowy plain dotted with pine trees, and the afterglow of the setting sun in the sky sprinkles a gentle glow. The video begins with the glass snowflake in the center of the frame, with sunlight passing through its transparent body, making it shine with colorful light. The snowflake's design is detailed, with clear edges and corners. The camera slowly rotates to the right around the snowflake, the distant pine trees are naturally distributed, appearing somewhat bent under the weight of the snow on the layered slopes. The camera continues to slowly rotate to the right and around the snowflake, another mountain view gradually comes into sight, with a few tall pine trees standing on the hilltop. On the horizon, the sun is about to set, and the remaining light turns the sky from light blue to warm orange. The camera continues to slowly rotate to the right around the snowflake, finally, the frame stays on the central glass snowflake, where the distant mountain top meets the horizon, and sunlight reflects on the snow.

Figure 11. **Snow example.** The videos generated by *DropletVideo*, Kling, and Vivago all maintain consistency with the prompt in terms of camera movement and various elements within the video. Their video quality is at the same level.



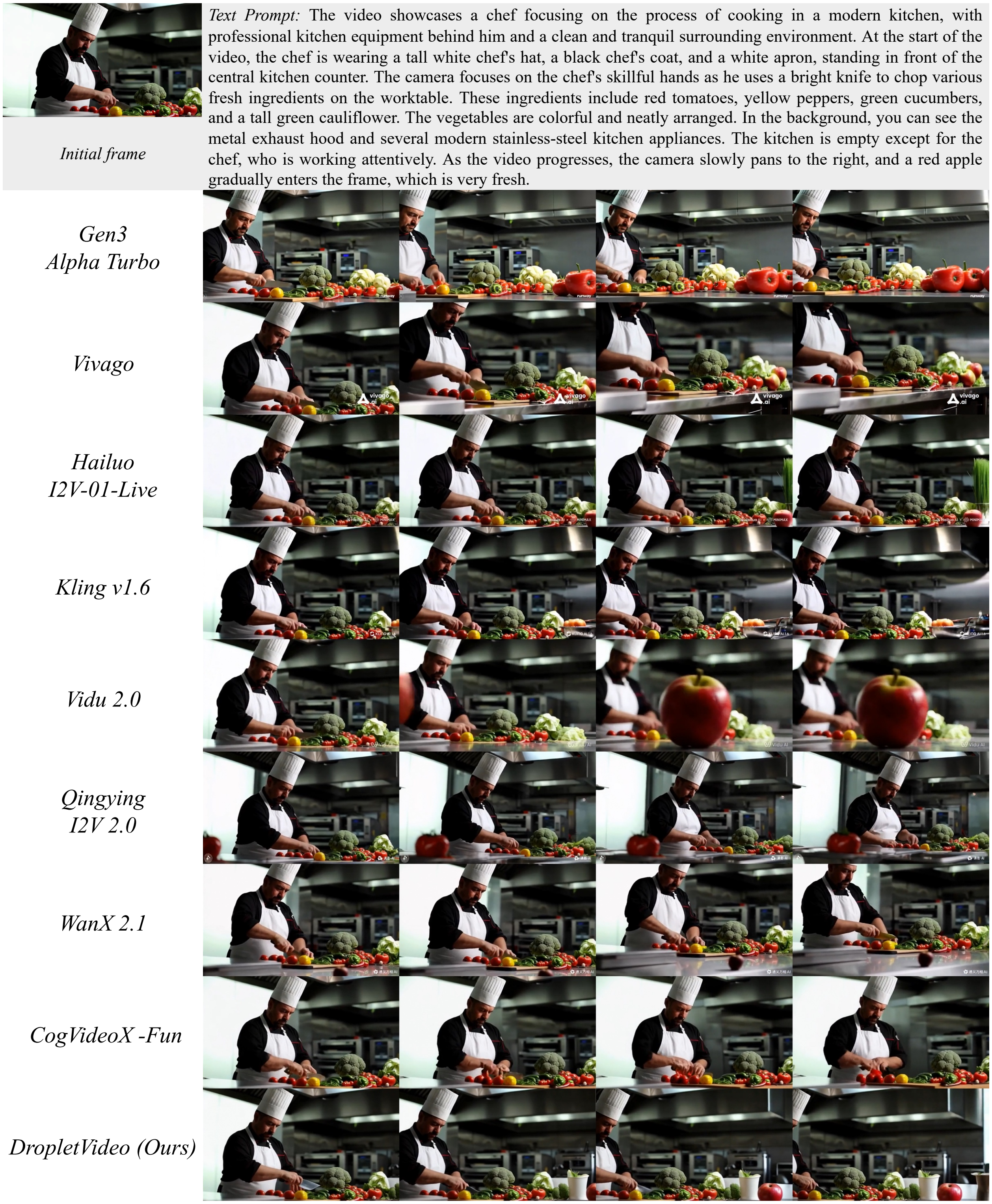
Text Prompt: The video shows a pair of small boats floating peacefully on a tranquil lake, with a magnificent sunset sky as the backdrop. the boat on the right is slowly chasing the boat on the left, with a soft golden glow reflecting the afterglow of the setting sun. The camera slowly moves from right to left, gradually revealing more background details. The distant city skyline appears hazy and dreamlike under the sunset, with a few tall buildings faintly visible. On the left side of the frame, tree branches sway gently in the breeze, adding a touch of natural movement to the scene. As the camera continues to move left, another small boat is shown quietly moored on the water on the left side, contrasting sharply with the distant city buildings.

Figure 12. **Boat example.** Our *DropletVideo*, along with Hailuo, WanX, and Kling v1.6, correctly understood the movement of the boat and the camera motion. However, these three models failed to ensure that the motion of the leaves remained logically consistent with the camera movement, resulting in the leaves moving synchronously with the camera, which is an unnatural effect. In contrast, our model maintains the relative motion consistency between the camera, boat, and leaves in the generated video. This is a typical demonstration of its integral spatio-temporal consistency capability.



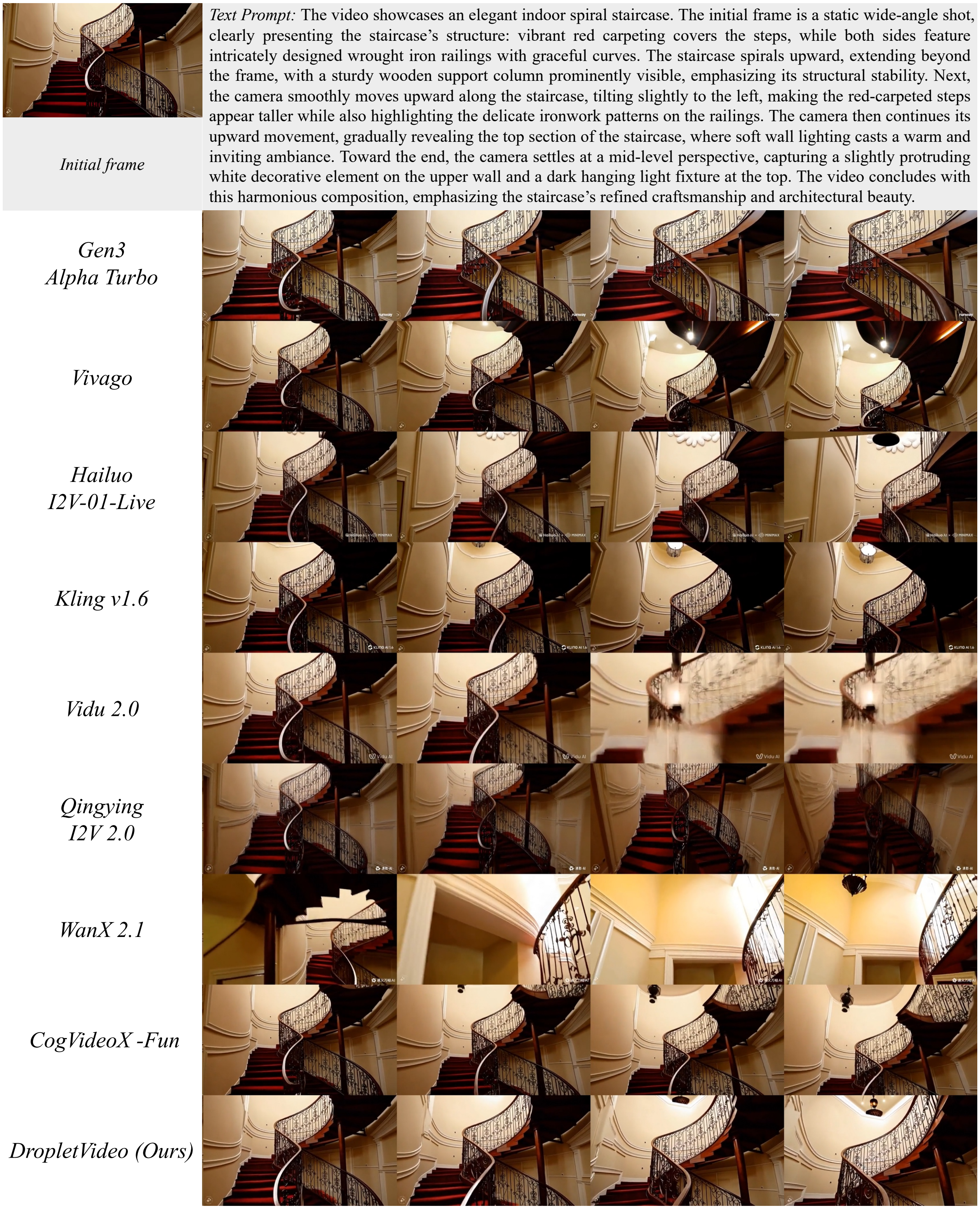
Text Prompt: The video presents a serene and beautiful sunset scene, capturing a flock of birds soaring gracefully under the evening sun, creating a stunning visual. The sun is slowly descending towards the horizon, painting the entire sky in warm shades of orange and red. The clouds, illuminated by the sunset, glow in golden hues, adding to the magnificent scenery. At the center of the frame stands a solitary tree, its branches appearing particularly distinct against the backdrop of the setting sun. As the camera moves slowly, a rolling grassland gradually emerges on the left side of the frame. The grassland, bathed in the sunset's afterglow, displays varying shades of light and shadow, adding a rhythmic natural beauty to the scene. As the camera continues to pan left, the flight path of the birds becomes increasingly visible, forming a bright arc under the glow of the sunset and enhancing the dynamic beauty of the composition. Further along, another tree appears in the frame, its silhouette sharply defined under the warm hues of the setting sun, with crisp and well-defined lines.

Figure 13. **Sunset example.** Only *DropletVideo* and Kling v1.6 successfully ensure the correct alignment between camera movement and object positioning. However, in Kling's generated video, the lighting reflections on the clouds remain unchanged, lacking natural variation. In contrast, in our model's generated video, as the camera moves, the light reflections on the clouds dynamically adjust, making the scene more consistent with real-world natural phenomena.



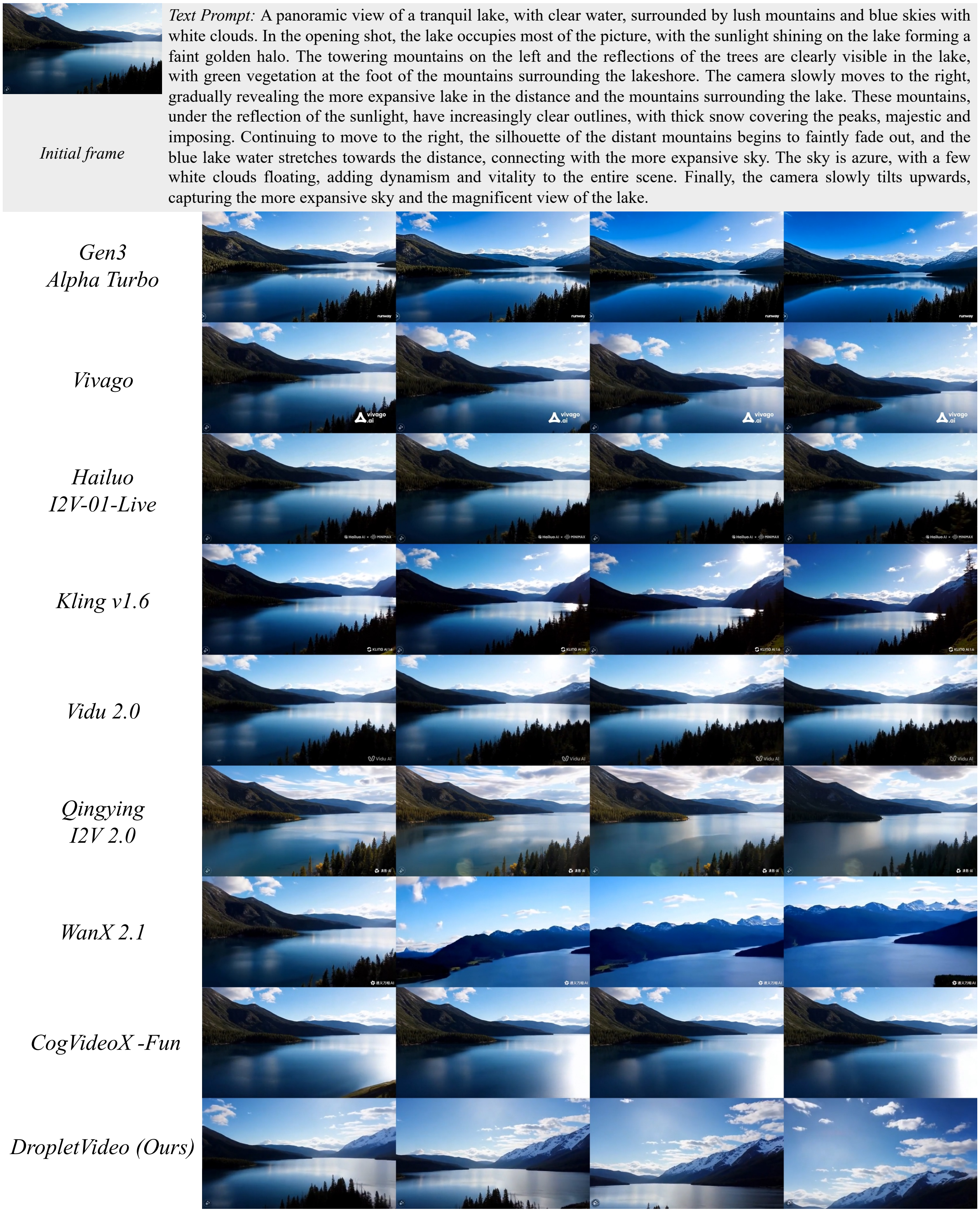
Text Prompt: The video showcases a chef focusing on the process of cooking in a modern kitchen, with professional kitchen equipment behind him and a clean and tranquil surrounding environment. At the start of the video, the chef is wearing a tall white chef's hat, a black chef's coat, and a white apron, standing in front of the central kitchen counter. The camera focuses on the chef's skillful hands as he uses a bright knife to chop various fresh ingredients on the worktable. These ingredients include red tomatoes, yellow peppers, green cucumbers, and a tall green cauliflower. The vegetables are colorful and neatly arranged. In the background, you can see the metal exhaust hood and several modern stainless-steel kitchen appliances. The kitchen is empty except for the chef, who is working attentively. As the video progresses, the camera slowly pans to the right, and a red apple gradually enters the frame, which is very fresh.

Figure 14. **Kitchen example.** We expect the focus of the video to transition from the chef to a red apple as the camera moves. Only *DropletVideo* successfully achieved this transition, while other models failed to correctly generate "a red apple" after the camera movement. Besides, it also ensures that the apple it generates are of a reasonable size and are positioned appropriately within the scene.



Text Prompt: The video showcases an elegant indoor spiral staircase. The initial frame is a static wide-angle shot, clearly presenting the staircase's structure: vibrant red carpeting covers the steps, while both sides feature intricately designed wrought iron railings with graceful curves. The staircase spirals upward, extending beyond the frame, with a sturdy wooden support column prominently visible, emphasizing its structural stability. Next, the camera smoothly moves upward along the staircase, tilting slightly to the left, making the red-carpeted steps appear taller while also highlighting the delicate ironwork patterns on the railings. The camera then continues its upward movement, gradually revealing the top section of the staircase, where soft wall lighting casts a warm and inviting ambiance. Toward the end, the camera settles at a mid-level perspective, capturing a slightly protruding white decorative element on the upper wall and a dark hanging light fixture at the top. The video concludes with this harmonious composition, emphasizing the staircase's refined craftsmanship and architectural beauty.

Figure 15. **Staircase example.** We required the camera to move smoothly up the stairs, ensuring that its trajectory remains logically consistent with the staircase in the video. Only our *DropletVideo* and Gen3 successfully maintained the correct camera movement path. However, Runway failed to generate key elements such as wall decorations and lights.



Text Prompt: A panoramic view of a tranquil lake, with clear water, surrounded by lush mountains and blue skies with white clouds. In the opening shot, the lake occupies most of the picture, with the sunlight shining on the lake forming a faint golden halo. The towering mountains on the left and the reflections of the trees are clearly visible in the lake, with green vegetation at the foot of the mountains surrounding the lakeshore. The camera slowly moves to the right, gradually revealing the more expansive lake in the distance and the mountains surrounding the lake. These mountains, under the reflection of the sunlight, have increasingly clear outlines, with thick snow covering the peaks, majestic and imposing. Continuing to move to the right, the silhouette of the distant mountains begins to faintly fade out, and the blue lake water stretches towards the distance, connecting with the more expansive sky. The sky is azure, with a few white clouds floating, adding dynamism and vitality to the entire scene. Finally, the camera slowly tilts upwards, capturing the more expansive sky and the magnificent view of the lake.

Figure 16. **Lake example.** The camera movement path is complex—it first moves to the right, then tilts upward, while the elements in the video change accordingly. All other models failed to accurately capture this camera movement, except for our *DropletVideo*. Our model not only strictly followed the prompt in executing the camera motion but also dynamically altered the scene, successfully revealing the sky and white clouds, which were not present in the initial image.

C.2. Quantitative Evaluation

C.2.1. Dense Prompt Rewrite

To effectively address the variability in language style and length of user-provided prompts, and to offer detailed guidance for video generation, we implement a dense prompt generation preprocessing step. This step serves as a bridge between the *DropletVideo* system and user input. Specifically, considering the superior performance of large language models in tasks such as text reasoning and image summarization, we have fine-tuned the InternVL2[27] model with instruction tuning. This fine-tuning is done using the LoRA[14], utilizing caption pairs from a high-quality training set. Experimental results indicate that approximately 600 such samples are sufficient to achieve the desired level of fine-tuning.

The module is designed to rephrase user prompts while keeping their original semantics intact. It transforms them into a standardized information architecture, akin to the trained captions. The module parses plot and camera movement details from the user input. It expands the content based on the input image, ensuring that the user's intent is preserved and detailed information is added. Furthermore, the module offers support for multiple languages.

We have revised 1,118 standard prompts supplied by VBench++ [15], resulting in the same number of comprehensive prompts, which we have labeled VBench++-ISTP (Integral Spatio-Temporal Prompts). These revised prompts incorporate both temporal and spatial variations. For instance, consider the original VBench++ prompt: *"A couple of horses are running in the dirt."* This has been rephrased to: *"The video showcases a dynamic scene of two horses running through mud, full of vitality and movement. The camera captures them kicking up dust, embodying a sense of freedom and abandon. The background faintly reveals the outlines of trees, adding a touch of natural tranquility to the entire scene. As the camera moves, the horses' running paths become clearer, and the dust sparkles in the sunlight, creating a dynamic visual effect."* Compared to the original prompt, the rephrased version provides a more detailed depiction as the camera moves, effectively introducing spatio-temporal information.

C.2.2. VBench++-IST Quantitative Results

We carried out an extensive evaluation of *DropletVideo*. For this purpose, we employed the evaluation code and the core performance metrics supplied by VBench++[15]. Furthermore, we integrated our integral spatio-temporal prompts, VBench++-ISTP, alongside the images from VBench++ [15]. In particular, we have refined all prompts to include comprehensive detail, as mentioned in Sec. C.2.1. In our comparative analysis, *DropletVideo* was benchmarked against the latest cutting-edge image-to-video models, including I2VGen-XL[33], Animate-Anything[19], and Nvidia-Cosmos[1].

Quantitative results are presented in Tab. 1. *DropletVideo* outperforms the other three models in most performance metrics. In terms of I2V Subject, I2V Background and Motion Smoothness, *DropletVideo*'s performance is 98.51%, 96.74%, and 98.94% respectively, both surpassing the other three models. In the aspect of Camera Motion, *DropletVideo* performs at 37.93%, significantly higher than the 12.95% of I2VGen-XL, the 10.64% of Animate-Anything, and the 37.56% of Nvidia-Cosmos. This suggests a strong capability of *DropletVideo* in handling camera motion within videos. For the Dynamic Degree, *DropletVideo*'s performance surpasses I2VGen-XL and Animate-Anything, yet falls below Nvidia-Cosmos, indicating a competitive performance of *DropletVideo* in maintaining motion coherence and dynamic degree.

In conclusion, despite some metrics where *DropletVideo* falls short compared to other models, it exhibits significant advantages in most of the key performance metrics. We believe that with further optimization and improvements, *DropletVideo* will be able to reach or even surpass the performance of other advanced models.

D. Future work

In future work, we will further investigate this issue by refining the data filtering strategies and expanding the dataset to a larger scale, with emphasis on diverse camera motions and dynamic objects. Furthermore, the types of camera motions supported by VBench++[15] are very limited, which is insufficient to capture the richness of spatial variations. It is worth exploring the development of a fine-grained camera motion classification model to better evaluate complex camera movements. Additionally, more suitable evaluation metrics should be proposed to comprehensively assess integral spatio-temporal consistency. Additionally, given the model's strong 3D consistency capability, we plan to extend its application to 3D/4D content generation.

Table 1. **Comparison of *DropletVideo* with state-of-the-art image-to-video models.** *DropletVideo* outperforms other models in **I2V Subject**, **I2V Background**, **Motion Smoothness** and **Camera Motion**. Meanwhile, *DropletVideo* remain at the current mainstream level for other metrics. In this table, **I2V-S** stands for I2V Subject, **I2V-B** stands for I2V Background, **CM** stands for Camera Motion, **SC** stands for Subject Consistency, **BC** stands for Background Consistency, **TF** stands for Temporal Flickering, **MS** stands for Motion Smoothness, **DD** stands for Dynamic Degree, **AQ** stands for Aesthetic Quality, **IQ** stands for Imaging Quality.

Models	**I2V-S**	**I2V-B**	**CM**	SC	BC	TF	**MS**	DD	AQ	IQ
I2VGen-XL[33]	96.08	94.67	12.95	95.76	97.67	97.40	98.27	24.80	65.26	69.21
Animate-Anything[19]	98.13	96.05	10.64	98.18	97.46	98.15	98.52	2.52	66.42	71.89
Nvidia-Cosmos[1]	95.10	95.30	37.56	91.59	94.43	96.20	98.82	83.90	58.39	70.35
DropletVideo (Ours)	**98.51**	**96.74**	**37.93**	96.54	97.02	97.68	**98.94**	27.97	60.94	70.35

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 18, 19

[2] Zhipu ai. qingying. https://chatglm.cn/video, 2024. 8

[3] AIGC-Apps. Cogvideox-fun. `https://github.com/aigc-apps/CogVideoX-Fun`, 2024. 8

[4] Staphord Bengesi, Hoda El-Sayed, Md Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers. *IEEE Access*, 2024. 6

[5] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024. 2, 3

[6] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 1, 4

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 2, 3

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2, 3

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis, 2024. *URL https://arxiv. org/abs/2403.03206*, 2. 6

[11] Hailuo. Hailuo ai. https://hailuoai.video, 2024. 8

[12] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 2, 3

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 6

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5, 18

[15] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 18

[16] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024. 1

[17] kuaishou. kuaishou-klingai. *https://klingai.kuaishou.com*, 2024. 8, 9

[18] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan. *apr*, 2024. 1

[19] Guojun Lei, Chi Wang, Hong Li, Rong Zhang, Yikai Wang, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. *arXiv preprint arXiv:2411.10836*, 2024. 18, 19

[20] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 2

[21] 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. *arXiv preprint arXiv:2407.02371*, 2024. 1

[22] Pexels. https://www.pexels.com. 1

[23] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10619–10629, 2022. 6

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 6

[25] Runway. Gen-3 alpha. https://runwayml.com/research/introducing-gen-3-alpha, 2024. 8

[26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[27] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. 18

[28] Tongyi. wanxiang. https://tongyi.aliyun.com/wanxiang/videoCreation, 2024. 8

[29] vidu ai. vidu. https://www.vidu.com, 2024. 8

[30] vivago ai. vivago. https://vivago.ai/video-generation, 2024. 8

[31] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20144–20154, 2023. 2

[32] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023. 6

[33] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 18, 19