

VLIPP: Towards Physically Plausible Video Generation with Vision and Language Informed Physical Prior

Supplementary Material

A. Coarse-Level Motion Planning Details

In this section, we present the experimental setting and details for reproducing the results. The main principle of our experimental setting is to fairly compare different Video Diffusion Models (VDMs) in generating physically plausible videos. Our adapt well-known open source model to serve as Compared Models. We discuss these models in details.

1. **CogVideoX[9]**: CogVideoX is capable of performing both text-to-video generation and image-to-video generation. It provides two model variants, featuring 2 billion and 5 billion parameters, respectively. In our experiments, we configured CogVideoX to generate 49 frames with a resolution of 720×480.
2. **LTX-Video[2]**: LTX-Video is also capable of performing both text-to-video generation and image-to-video generation. In our experiments, we compared two versions of LTX-Video with corresponding methods. It can generate videos with 49 frames with a resolution of 768×512.
3. **SVD-XT[1]**: SVD-XT is capable of performing image-to-video generation. In our experiments, we configured SVD-XT to generate 25 frames with a resolution of 1024×576.
4. **SG-I2V[8]**: SG-I2V is a motion trajectory-guided image-to-video generation model. It is capable of generating bounding box-controllable videos with 14 frames with a resolution of 1024×576.
5. **LLM-grounded Video Diffusion Models[5]**: LLM-grounded Video Diffusion Models are capable of predicting future frame bounding boxes based on input prompts and injecting the box information in a training-free manner. In our experiments, we configured LLM-grounded to generate 24 frames with a resolution of 576×320.

We additionally present the Reasoning Template utilized during the stage 1 Coarse-Level Motion Planning process, as shown in Fig 1 and Fig 2. This includes system instructions to ensure the proper functioning of the chain of thought and provides the VLM with context information to guarantee the accuracy of predictions.

B. Experiment Details

In this section, we present the experimental details of our benchmark, PhyGenBench[6] and Physics-IQ[7].

PhyGenBench comprises 160 prompts, spanning four

domains of physical knowledge: Mechanics (40), Optics (50), Thermal (40), and Material (20), along with 27 types of physical laws. It also includes 165 objects and 42 actions. The evaluation focuses on two aspects: semantic alignment and physical commonsense alignment. The degree of semantic alignment is assessed by extracting objects and actions from the prompts using a Vision-Language Model (VLM), determining whether the objects appear, and evaluating based on the presence of objects and the occurrence of actions. The degree of physical commonsense alignment is determined through a three-step process: detecting whether the physical phenomena occur and whether the order of occurrence is correct; and finally conducting an overall naturalness evaluation.

Physic-IQ categorizes real-world physical laws into Solid Mechanics, Fluid Dynamics, Optics, Magnetism, and Thermodynamics, encompassing 114, 45, 24, 6, and 9 videos, respectively. The evaluation approach is twofold, focusing on physical comprehension and visual authenticity. Physical comprehension is determined by identifying the timing, location, and frequency of actions, ultimately calculating the mean squared error between corresponding pixels in the generated and real frames to derive a physical comprehension score. Visual authenticity is evaluated using a Vision-Language Model (VLM), employing the gold standard experimental method from psychophysics. The VLM receives pairs of real and generated videos of the same scene in random order and is tasked with identifying the real scene, a design intended to reflect visual authenticity.

During the experimental phase of this paper, we utilized the prompts provided by the PhyGenBench dataset to infer the initial frame’s prompts using an LLM, which were then generated by FLUX[3]. To ensure fairness in comparison, all I2V models were supplied with the same initial frame image. Given that different models produce videos with varying numbers of frames, a uniform sampling ratio was applied during the testing phase to extract key frames consistently across all models.

C. User Study

To complement our above evaluations, we conduct a user study to assess the subjective human perception of the generated videos. We follow the gold standard experimental approaches from psychophysics, a 2AFC paradigm, which means two-alternative-forced-choice [7]. In our case, participants completed a questionnaire in which they were presented with pairs of videos and asked to select the one that

better aligned with their expectations of physical realism. Responses from 50 participants are summarized in Table 1. The result indicates a strong preference for videos generated by our framework over those from competitors. A detailed analysis of these findings follows in the subsequent discussion.

Model	P.P.(↑)	V.R.(↑)
CogVideoX-I2V-5B	34%	40%
LTX-Video	22%	18%
Ours	52%	48%

Table 1. Ablation study on input condition and Motion Animation Module.

D. Additional Ablation Study

To validate the effectiveness of the key components in VLIPP, we conduct two additional ablation experiments. First off, we replaced the optical flow condition with 2D bounding boxes condition. We adopt MagicMotion [4], a CogVideoX variant for 2D bounding boxes conditioned generation. Moreover, we disable the Motion Animation Module and extract optical flow directly from the 2D bounding boxes. Table 2 rows 1&3 shows that optical flow works as a more effective motion condition for VDMs than 2D bounding boxes control. Comparing Table 2 rows 2&3 show that optical flow from Motion Animation Module outperforms optical flow derived from 2D bounding boxes.

Method	Solid ↑	Fluid ↑	Optics ↑	Magnet ↑	Ther ↑	Avg ↑
MagicMotion	34.9	30.2	16.7	13.0	8.5	29.8
Ours(w/o simu)	38.7	34.1	16.6	13.2	7.7	32.8
Ours	42.3	34.1	16.9	13.4	8.8	34.9

Table 2. User study statistics of the preference rate for Physical Plausibility (P.P.) & Visual Realism (V.R.).

E. More Qualitative Results

In this section, we further demonstrate examples of the proposed framework across various scenarios using prompts from the PhyGenBench dataset. Fig 3, 4, and 5 show video generation results driven by physical conditions like fluid dynamics and thermodynamics, while Fig 6 highlights effects in light reflection scenarios.

References

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1
- [2] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Real-time video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 1
- [3] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 1
- [4] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *arXiv preprint arXiv:2503.16421*, 2025. 2
- [5] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*, 2023. 1
- [6] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 1
- [7] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025. 1
- [8] Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, and David B Lindell. Sg-i2v: Self-guided trajectory control in image-to-video generation. *arXiv preprint arXiv:2411.04989*, 2024. 1
- [9] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiao-han Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1


Physics-Aware Reasoning Template (Gravity/Mouementum)

System Instruction


You are an expert in real-world physical motion. Your task is to predict video bounding boxes. You will receive a reference image, its segmentation map, the initial bounding box where motion starts, and a textual description of the video content. Your task is to predict the bounding box coordinates for the next 13 frames. When predicting, ensure that different bounding boxes do not overlap or exceed the frame boundaries.

Each frame should be represented as: '[{"id": unique object identifier, "name": object name, "box": [box top-left x-coordinate, box top-left y-coordinate, box width, box height]}, ...]'. Your predicted frames must cover the entire described scene. Critical moments in the movement must be included in the predicted sequence, such as the moment of collision. You need to consider the material and properties of the moving objects in the video. You only need to focus on the objects provided to you in the first frame, and no need to pay attention to anything else. Your predictions should infer which physical laws apply, such as Gravity, Conservation of Momentum, Fluid Mechanics, Thermodynamics, Magnetic Force, etc. For example, under the action of gravity, the ball will accelerate and rebound after colliding with the ground. When two balls move towards each other, the material, volume, density, etc. of each ball must be considered, and the collision process satisfies the law of conservation of momentum. Assume objects move and interact based on real-world physics. The camera follows perspective geometry. Even if an object disappears and reappears, its bounding box should maintain the same ID across frames. If necessary, you may make reasonable assumptions. Before generating bounding box coordinates for each frame, provide step-by-step reasoning.

Context Information

 **User:** Caption: Generate a high-quality video of a wooden ball dropping from a height of about 1 meter onto a smooth, solid table. Capture the physics of the ball's descent, including its acceleration due to gravity, the moment of impact with the table, and the subsequent bounce. Show the ball compressing slightly upon impact, then rebounding, losing a bit of height with each bounce until it comes to rest. Include subtle details like the sound of the impact, the slight vibration of the table, and the ball's rotation as it moves.

initial_boxes: [{"id": 0, "name": "ball", "box": [303.0, 82.2, 108, 108]}]

 **Assistant:** Reasoning:

Step1-Analysis Caption and determine the physical laws to be satisfied: As can be seen from the caption, the topic of this study is about the free fall of a small ball, so this topic needs to satisfy Newton's laws of motion, focusing on the impact of gravity.

Step2-Analyze the impact of this physical law: In accordance with Newton's laws of motion, gravity will cause the small ball to perform free fall motion, vertically accelerate and hit the table, and rebound after the collision. Due to energy loss, the height of each rebound will decrease, which is in line with the law of conservation of energy and momentum.

Step3-Analyze the impact on box coordinates: The horizontal motion of the ball is negligible, so the x coordinate remains basically unchanged. Due to the vertical acceleration, the y coordinate of the small ball will gradually increase and the interpolation of the y coordinate between each two frames will also gradually increase. After colliding with the table, the small ball rebounds, the y coordinate decreases, and the small ball decelerates vertically upward. Subsequently, the small ball continues to accelerate vertically and collides with the table for the second time. After the third collision, the small ball is stationary on the table, and the y coordinate remains unchanged. Since the ball is a rigid body, it will be slightly compressed at the moment of collision with the desktop. Therefore, the width in the box coordinates remains unchanged, the height decreases slightly, and then quickly returns to normal. The width and height will not change for the rest of the time.

Predict the coordinates for the next 13 frames:

```
Frame 1: [{"id": 0, "name": "wooden ball", "box": [303, 82, 108, 108]}]
Frame 2: [{"id": 0, "name": "wooden ball", "box": [303, 137, 108, 108]}]
Frame 3: [{"id": 0, "name": "wooden ball", "box": [303, 184, 108, 108]}]
Frame 4: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 5: [{"id": 0, "name": "wooden ball", "box": [303, 255, 108, 108]}]
Frame 6: [{"id": 0, "name": "wooden ball", "box": [303, 284, 108, 104]}]
Frame 7: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 8: [{"id": 0, "name": "wooden ball", "box": [303, 298, 108, 108]}]
Frame 9: [{"id": 0, "name": "wooden ball", "box": [303, 305, 108, 108]}]
Frame 10: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 104]}]
Frame 11: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 12: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 108]}]
Frame 13: [{"id": 0, "name": "wooden ball", "box": [303, 313, 108, 104]}]
```

Figure 1. Physics-Aware Reasoning Template for Rigid Body Motion


Physics-Aware Recognition Template (Fluid)

System Instruction


You are an expert in real-world physical motion. Your task is to predict video bounding boxes. You will receive a reference image, its segmentation map, the initial bounding box where motion starts, and a textual description of the video content. Your task is to predict the bounding box coordinates for the next 13 frames. When predicting, ensure that different bounding boxes do not overlap or exceed the frame boundaries.

Each frame should be represented as: '[{"id": unique object identifier, "name": object name, "box": [box top-left x-coordinate, box top-left y-coordinate, box width, box height]}, ...]'. Your predicted frames must cover the entire described scene. Critical moments in the movement must be included in the predicted sequence, such as the moment of collision. You need to consider the material and properties of the moving objects in the video. Your predictions should infer which physical laws apply, such as Gravity, Conservation of Momentum, Fluid Mechanics, Thermodynamics, Magnetic Force, etc. For example, under the influence of fluid mechanics, when pouring water into a cup, the water level in the cup should rise and ripples should occur. The top boundary of a liquid should move when it's being poured, while the bottom edge of the liquid (for instance, in a container) should stay fixed. When an object (such as a liquid) is in a container and is being poured, the bottom boundary of the liquid typically stays fixed. Assume objects move and interact based on real-world physics. The camera follows perspective geometry. Even if an object disappears and reappears, its bounding box should maintain the same ID across frames. If necessary, you may make reasonable assumptions. Before generating bounding box coordinates for each frame, provide step-by-step reasoning.

Context Information

 **User:** Caption: The video shows a close-up of a clear glass being filled with tea. The tea is being poured from above, and we can see the stream of tea hitting the bottom of the glass and causing ripples and splashes. The background is a plain white surface, which contrasts with the transparency of the glass and the clarity of the tea. The glass is cylindrical in shape and appears to be of a standard size for a drinking glass.

initial_boxes: [{"id": 0, "name": "tea", "box": [241.8, 335.3, 165, 88]}]

 **Assistant:** Reasoning:

Step1-Analysis Caption and determine the physical laws to be satisfied: As can be seen from the caption, the theme of this session is to pour tea into a glass teacup. Therefore, the tea is affected by gravity during its fall, and the state of the tea in the cup satisfies fluid mechanics.

Step2-Analyze the impact of this physical law: Tea falls into the cup under the influence of gravity. According to fluid mechanics, the water level in the cup will gradually rise, and the water surface will ripple due to the pouring of water.

Step3-Analyze the impact on box coordinates: When tea is poured into the cup, under the restriction of the cup, the x coordinate of the tea box will not change much. As the water level rises, the y coordinate will gradually decrease. As the water level rises, the overall volume of the tea gradually increases, the height of the box will also increase, and the width will change with the boundary of the cup.

Predict the coordinates for the next 13 frames:

```
Frame 1: [{"id": 0, "name": "tea", "box": [241.8, 335.3, 165, 88]}]
Frame 2: [{"id": 0, "name": "tea", "box": [241.2, 330.0, 165, 94]}]
Frame 3: [{"id": 0, "name": "tea", "box": [241.0, 322.2, 165, 102]}]
Frame 4: [{"id": 0, "name": "tea", "box": [241.3, 317.6, 165, 105]}]
Frame 5: [{"id": 0, "name": "tea", "box": [240.6, 313.7, 166, 111]}]
Frame 6: [{"id": 0, "name": "tea", "box": [239.1, 305.6, 167, 118]}]
Frame 7: [{"id": 0, "name": "tea", "box": [237.7, 301.1, 170, 123]}]
Frame 8: [{"id": 0, "name": "tea", "box": [236.8, 295.3, 172, 129]}]
Frame 9: [{"id": 0, "name": "tea", "box": [235.7, 290.2, 172, 134]}]
Frame 10: [{"id": 0, "name": "tea", "box": [235.1, 283.3, 173, 140]}]
Frame 11: [{"id": 0, "name": "tea", "box": [234.0, 278.3, 175, 145]}]
Frame 12: [{"id": 0, "name": "tea", "box": [233.1, 272.0, 177, 152]}]
Frame 13: [{"id": 0, "name": "tea", "box": [231.7, 268.0, 179, 156]}]
```

Figure 2. Physics-Aware Reasoning Template for fluid dynamics and thermodynamics.



Figure 3. More examples of generated videos related to fluid dynamics and thermodynamics.

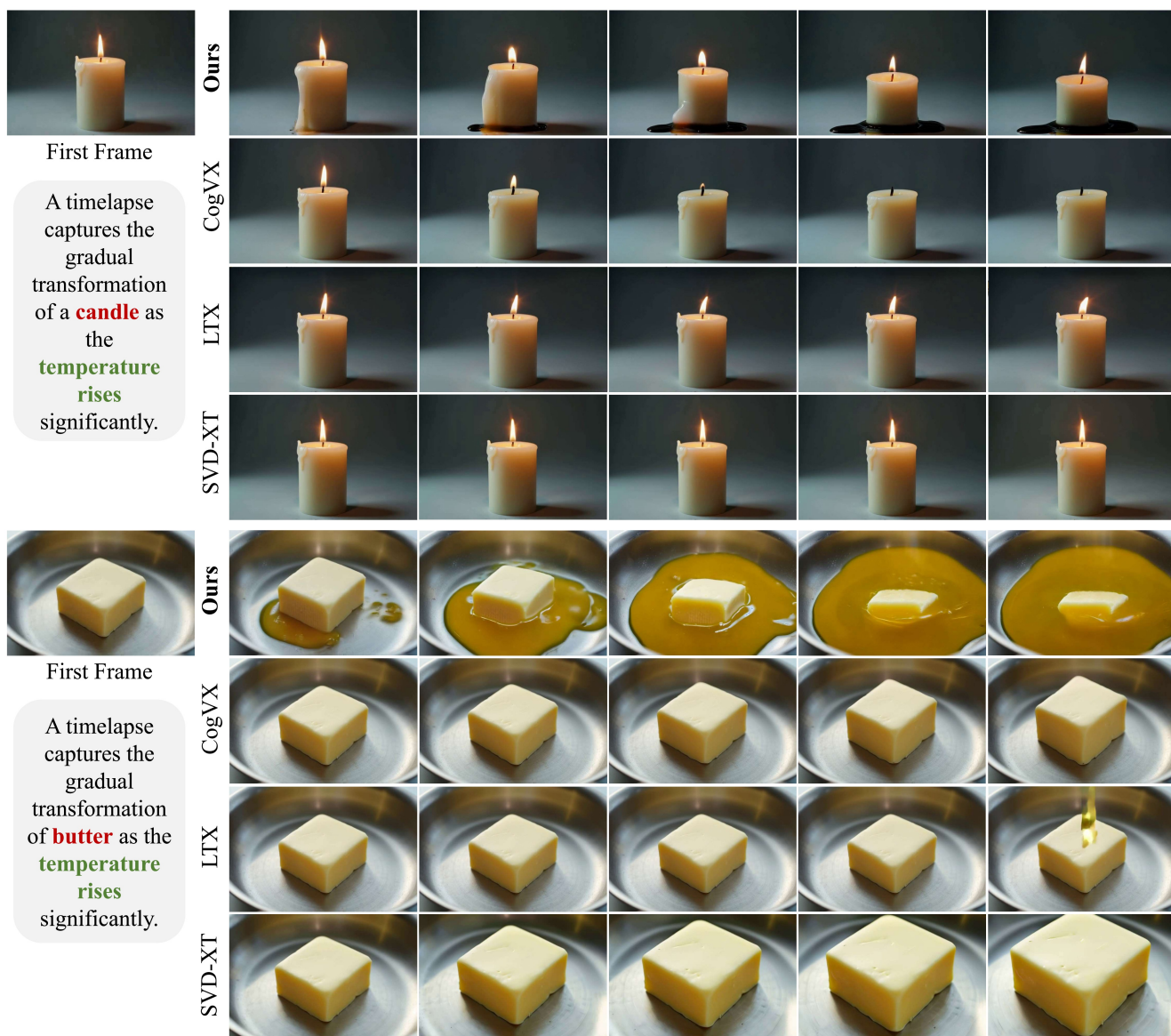


Figure 4. More examples of generated videos related to thermodynamics.

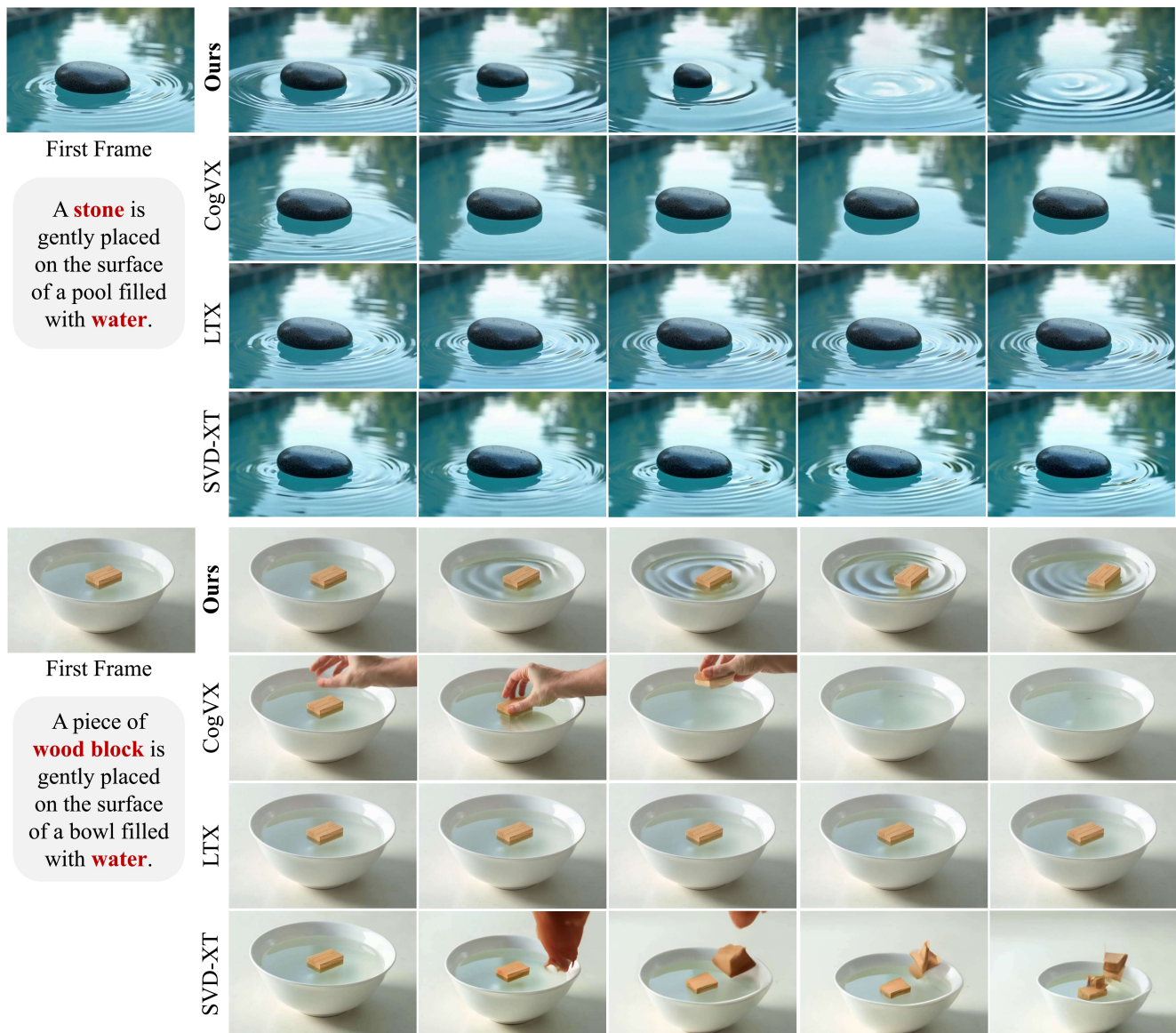


Figure 5. More examples of generated videos related to fluid dynamics.

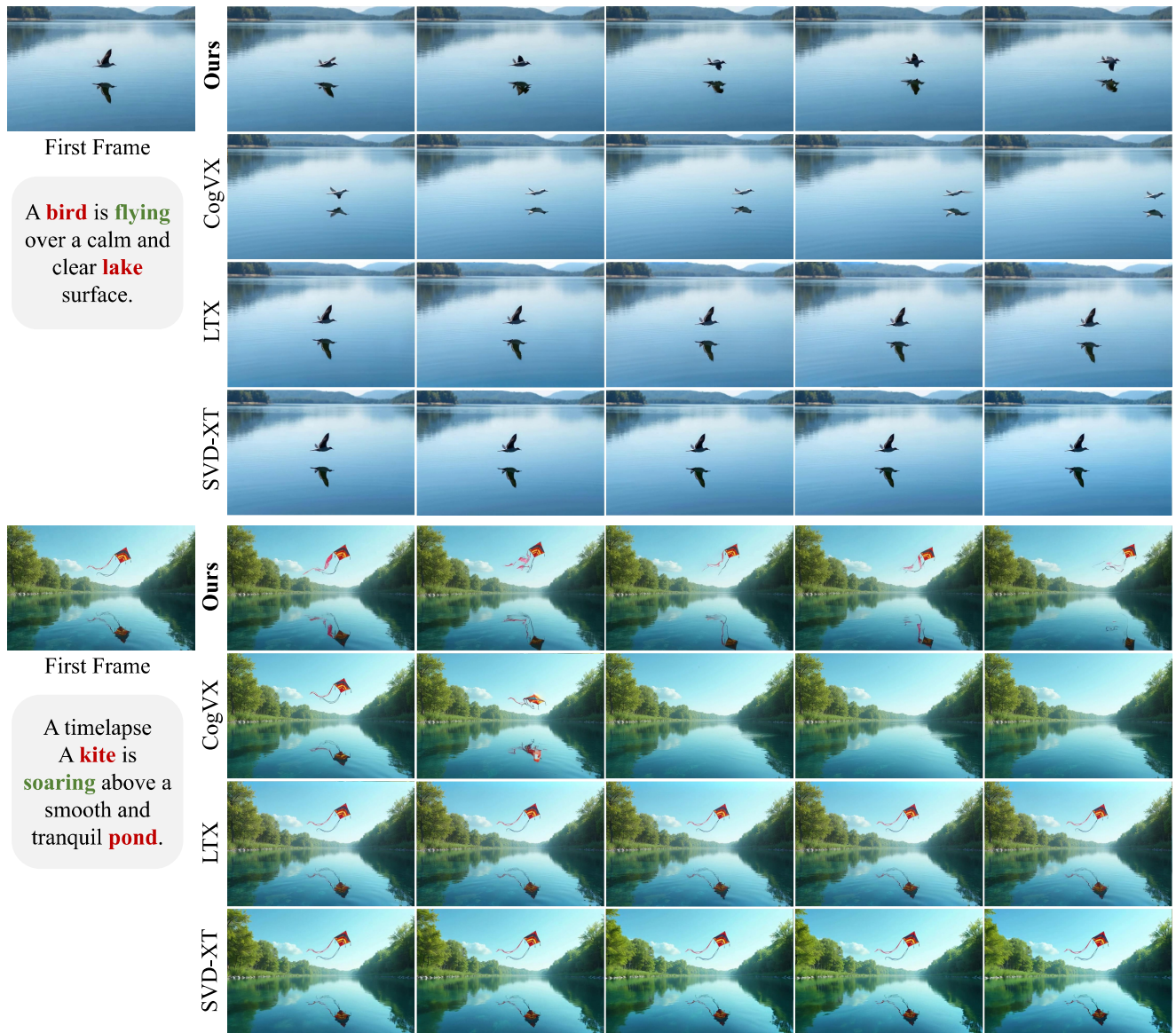


Figure 6. More examples of generated videos related to optics.