# Appendix

**Roadmap.** Section A introduces the ten baseline text-to-video models' implementation details. In Section B, we present the evaluation prompt template. In Section C, we display a range of video examples.

## A. Implementation Details

This section provides further details on the 10 baseline text-to-video models, as listed below:

- **Sora** [69]: Developed by the OpenAI team in 2024, Sora is a closed-source generative model. The model can generate 30 FPS videos with selectable durations of 5, 10, 15, or 20 seconds. It supports a range of output formats, including resolution from 480p to 1080p and multiple aspect ratios (1:1, 16:9, and 9:16). Sora provides style presets and can produce four distinct video variants from a single prompt. Additionally, a "relaxed mode" is available with a processing latency of approximately 30 seconds per video.
- **Dreamina Video 3.0** [8]: Dreamina Video 3.0 is a closed-source generator released by the Bytedance team in 2024, supporting both 5- and 10-second videos. It supports a wide range of aspect ratios (16:9, 21:9, 4:3, 1:1, 3:4, 9:16) and utilizes DeepSeek-R1 [29] for prompt enhancement.
- **Qingying** [98]: Qingying is the commercial implementation of Zhipu's open-source CogVideo models [39, 90]. It generates 5-second videos at 30/60 FPS across five aspect ratios of 1:1, 9:16, 16:9, 4:3, and 3:4. Qingying supports two modes: Quality and Fast. Additionally, it provides fine-grained control over video style, emotional atmosphere, and camera movement, alongside support for AI-generated audio and visual effects.
- **Wan2.1 Plus** [2]: Wan2.1 Plus is an open-source generative model [86] released by Alibaba Group in 2025, supporting multiple aspect ratios (1:1, 3:4, 4:3, 9:16, 16:9). It provides additional features such as "Inspiration Mode" and "Sound Effects".
- **Mochi-1** [24]: Released by Genmo in 2024, Mochi-1 is an open-source model. Its standard output consists of 5-second, 24 FPS video at 480p resolution with a 16:9 aspect ratio. Mochi-1 supports a seed function for reproducibility and includes a feature for random prompt suggestions. It can generate two videos simultaneously, with an approximate processing time of 3 minutes per video.
- **LTX Video** [35]: Developed by Lightricks in 2024, LTX Video is an open-source model. It generates 5-second, 24 FPS videos at 512p resolution, supporting 16:9, 1:1, and 9:16 aspect ratios. LTX Video enables fine-grained control over the location, shot type, references, and style, and even supports voiceover integration.
- **PixVerse V4.5** [1]: PixVerse V4.5 is a closed-source model from AISphere, released in 2025. It generates videos with a duration of either 5 or 8 seconds. PixVerse V4.5 supports multiple resolutions including 360p, 540p, 720p, and 1080p, and offers five aspect ratios: 16:9, 4:3, 1:1, 3:4, and 9:16.
- **Kling 1.6** [51]: Released by Kuaishou in 2024, Kling 1.6 is a closed-source generative model. It generates video outputs of 5 or 10 seconds in duration, supporting 16:9, 1:1, and 9:16 aspect ratios. It features two generation modes: a standard mode and a restricted high-quality mode. Kling supports advanced prompting functionalities, including negative prompts, fixed seeds for reproducibility, a prompt dictionary, and AI-assisted prompt suggestions. For generations, Kling can create 4 videos simultaneously from a single prompt. The processing time is approximately 4 minutes per video, with a maximum batch size of 5 videos.
- **Hailuo 01-Director** [64]: Hailuo 01-Director is a closed-source model released by Minimax in 2025 for text-to-video generation. Its standard output is a 6-second, 24 FPS video at 720p resolution, typically with a default aspect ratio of 16:9.
- **Pika 2.2** [71]: Pika2.2 is a closed-source generative model from Pika Labs, released in 2025. It provides Pikawaps, Pikaaddition, Pikaaffects, Pikaframes, and Pikascenes. Pika 2.2 generates 5- or 10-second videos at resolutions of 720p or 1080p, supporting a wide range of aspect ratios (16:9, 9:16, 1:1, 4:5, 4:3, 5:2). For generation control, it supports both negative prompts and seed inputs. Pika 2.2 can produce 4 videos simultaneously, with each taking approximately 30 seconds to process.

## B. Evaluation Prompt

We employ a prompt-based framework to conduct a comprehensive evaluation of AI-generated videos using the LLaVA model. This methodology involves combining a base prompt (see Figure 3) with a specialized prompt tailored to a specific evaluation dimension. For instance, evaluating video quality is achieved by pairing the Base Prompt with the Quality Prompt (see Figure 5). Similarly, video realism is evaluated using the Base Prompt plus the Realism Prompt (see Figure 6). The same approach is applied to evaluate video relevance using the Relevance Prompt (see Figure 7) and consistency using the

Consistency Prompt (see Figure 8). This ensures that each evaluation is grounded in a consistent context while allowing for a focused, independent score for each distinct attribute of the video.

YOUR TASK: Evaluate the TECHNICAL QUALITY of these consecutive frames.

Quality (1-5): Technical Excellence
Check for artifacts, resolution, clarity, color balance, and rendering quality across all consecutive frames.
- 1: Severe technical issues affecting most frames
- 2: Multiple obvious flaws impacting viewing experience
- 3: Acceptable with minor flaws
- 4: High quality with trivial imperfections
- 5: Flawless professional-grade

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of technical quality]
Quality: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The frames show good resolution and color balance with minor compression artifacts.
Quality: 4

Figure 5. **Template for Quality Prompt.** Quality Prompt is designed specifically for evaluating the video quality. It directs the AI evaluator to assess objective attributes such as artifacts, resolution, clarity, and color balance, using a 1-to-5 score to quantify the video quality from severely flawed to professional-grade.

## C. Video Examples

In this Section, we provide extensive examples of videos generated by our proposed benchmark prompts. Figure 9 and Figure 10 present the results of our quality study. Figure 11-20 shows the generation result of each text-to-video model in our benchmark, where five representative frames from each video are extracted and arranged sequentially to form a visual strip. These presented instances are consistent with the experimental setting discussed in Section 4

YOUR TASK: Evaluate the REALISM and believability of the content.

Realism (1-5): Physical Plausibility
Assess believability and natural appearance throughout the consecutive sequence.
- 1: Severe physics violations, obviously fake appearance
- 2: Multiple unnatural elements, clearly AI-generated look
- 3: Generally plausible with some artificial aspects
- 4: Very natural with minimal artificial tells
- 5: Photorealistic perfection

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of realism and believability]
Realism: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The video appears very natural with realistic physics and minimal artificial elements.
Realism: 4

Figure 6. **Template for Realism Prompt.** Realism Prompt guides the evaluation of the video realism. The assessment is based on physical plausibility, instructing the evaluator to identify any physics violations, unnatural elements, or other artificial tell. The 1-to-5 score quantifies how closely the content approximates photorealistic perfection.

YOUR TASK: Evaluate how well the frames match the generation goals.

Relevance (1-5): Adherence to Goals
Compare the video sequence against the Prompt: `{original_prompt}` and Explanation: `{explanation}`.
- 1: Completely unrelated content
- 2: Weak connection, missing major elements
- 3: Captures general concept, lacks details
- 4: Accurately represents most elements
- 5: Perfect alignment with all requirements

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of how well it matches the prompt and explanation]
Relevance: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The video accurately depicts most elements from the prompt but lacks some specific details.
Relevance: 4

Figure 7. **Template for Relevance Prompt.** Relevance Prompt focuses on evaluating video relevance to the user's prompt and explanation. It instructs the AI evaluator to compare the visual output against the provided original prompt and explanation, scoring the alignment on a 1-to-5 scale based on how accurately the generated content captures the required elements and world knowledge.

YOUR TASK: Evaluate the TEMPORAL CONSISTENCY between consecutive frames.

Consistency (1-5): Temporal Coherence Between Consecutive Frames
IMPORTANT: Since these are consecutive frames, analyze smooth transitions and logical progression from frame to frame.
- 1: Chaotic inconsistency - objects teleport, backgrounds change randomly, no logical flow between consecutive frames
- 2: Major temporal disruptions - significant jumps or morphing between adjacent frames, jarring transitions
- 3: Generally stable progression with some noticeable but minor temporal inconsistencies between frames
- 4: Smooth temporal flow with natural progression, only very minor variations between consecutive frames
- 5: Perfect temporal continuity - seamless, natural progression that could be from real video footage

CRITICAL OUTPUT FORMAT REQUIREMENT:
You MUST end your response with exactly this format (no variations allowed):

Reasoning: [Your detailed analysis of frame-to-frame consistency and temporal flow]
Consistency: X

Where X is ONLY a single digit from 1 to 5. Do not add brackets, extra text, or explanations after the number.

EXAMPLE:
Reasoning: The consecutive frames show smooth transitions with natural progression and minimal temporal inconsistencies.
Consistency: 4

Figure 8. **Template for Consistency Prompt.** Consistency Prompt is used to evaluate the video consistency. The core task is to analyze the coherence between consecutive frames, focusing on the smoothness of transitions and the logical progression of objects and actions. The 1-to-5 score measures the video's temporal flow, from chaotic and disjointed to seamless and natural.

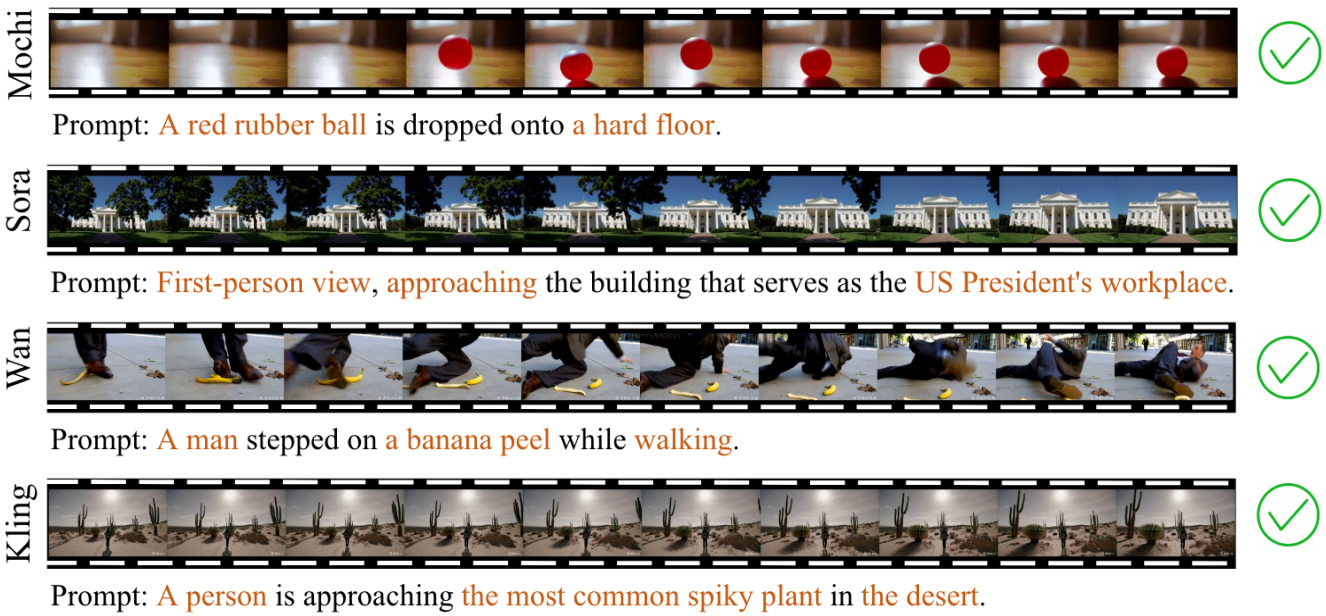


Figure 9. Examples of successfully understanding world knowledge.

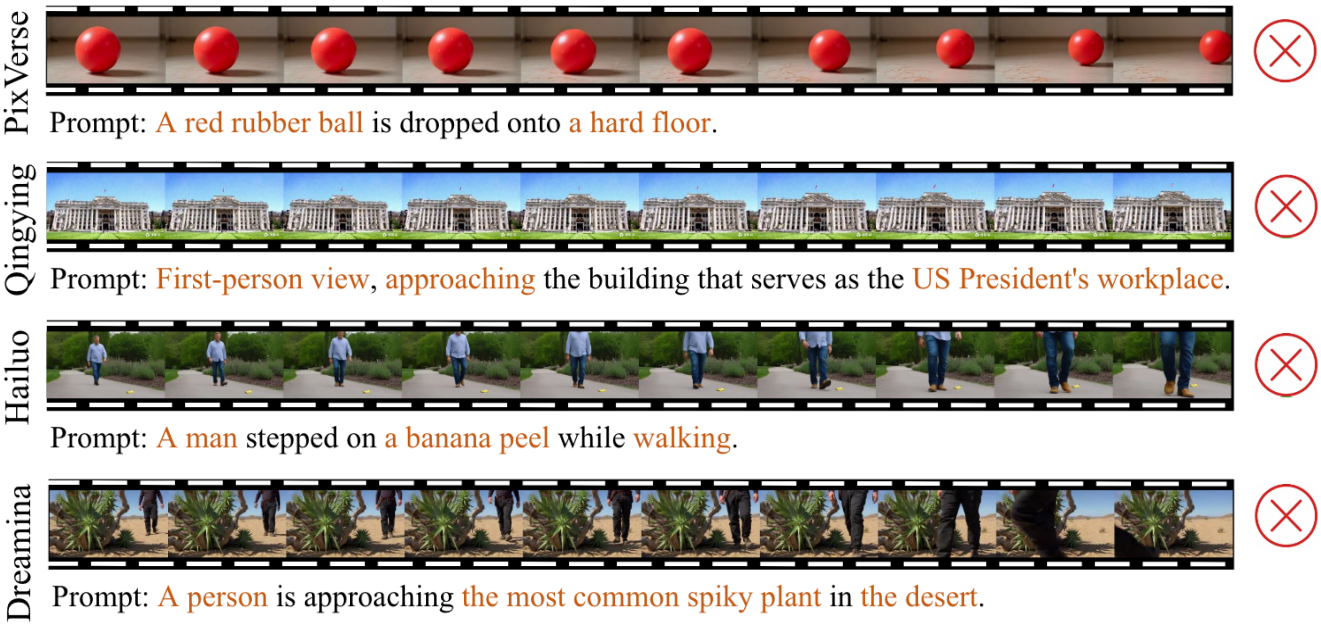


Figure 10. Examples of failures in understanding world knowledge.

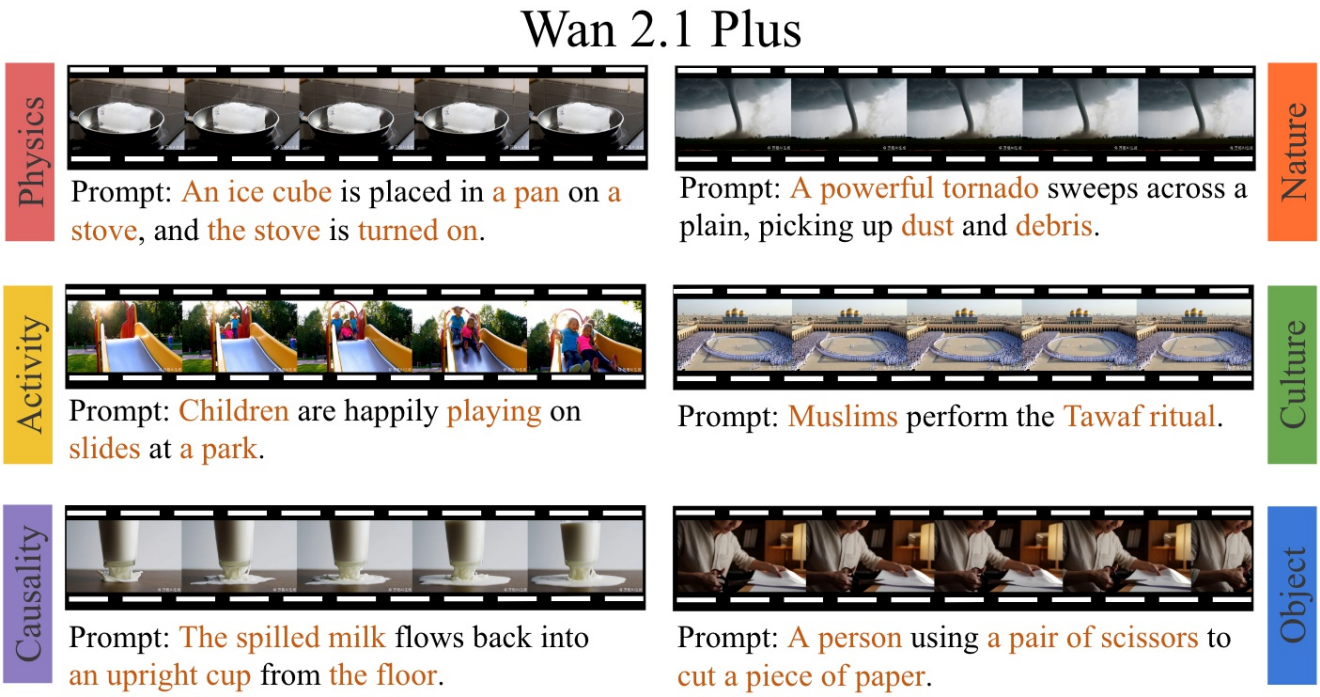


Figure 11. Video generation of Wan 2.1 Plus.

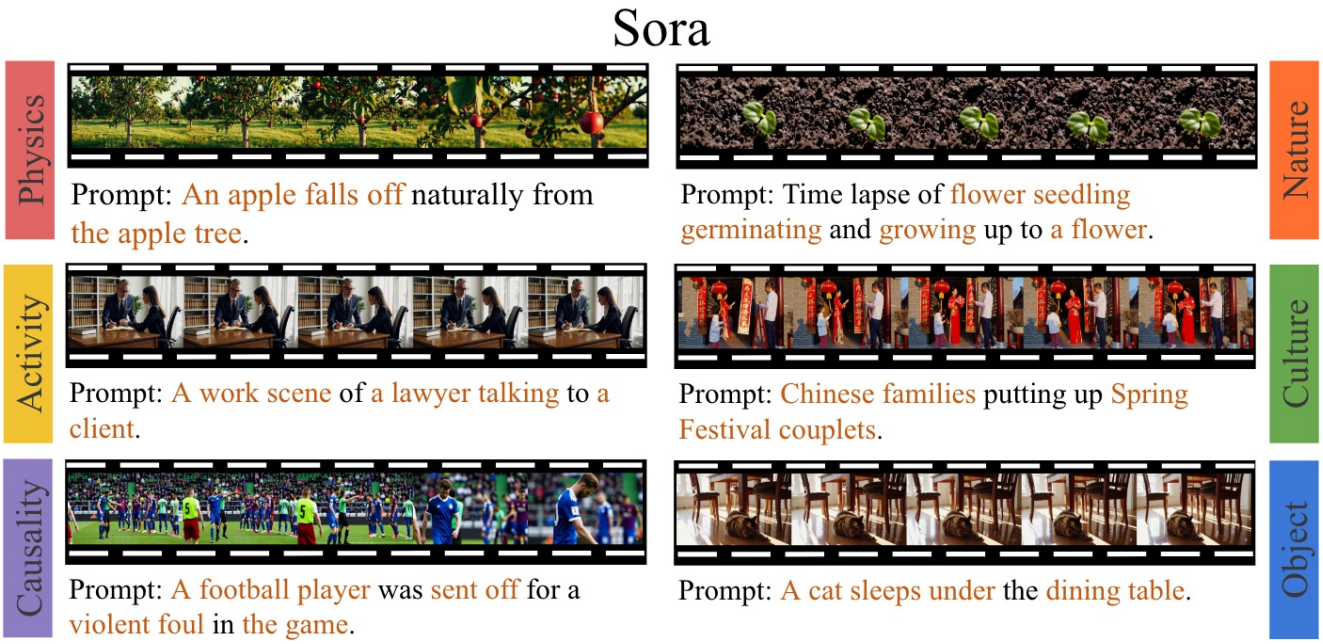


Figure 12. Video generation of Sora.


Figure 13. Video generation of Kling 1.6.

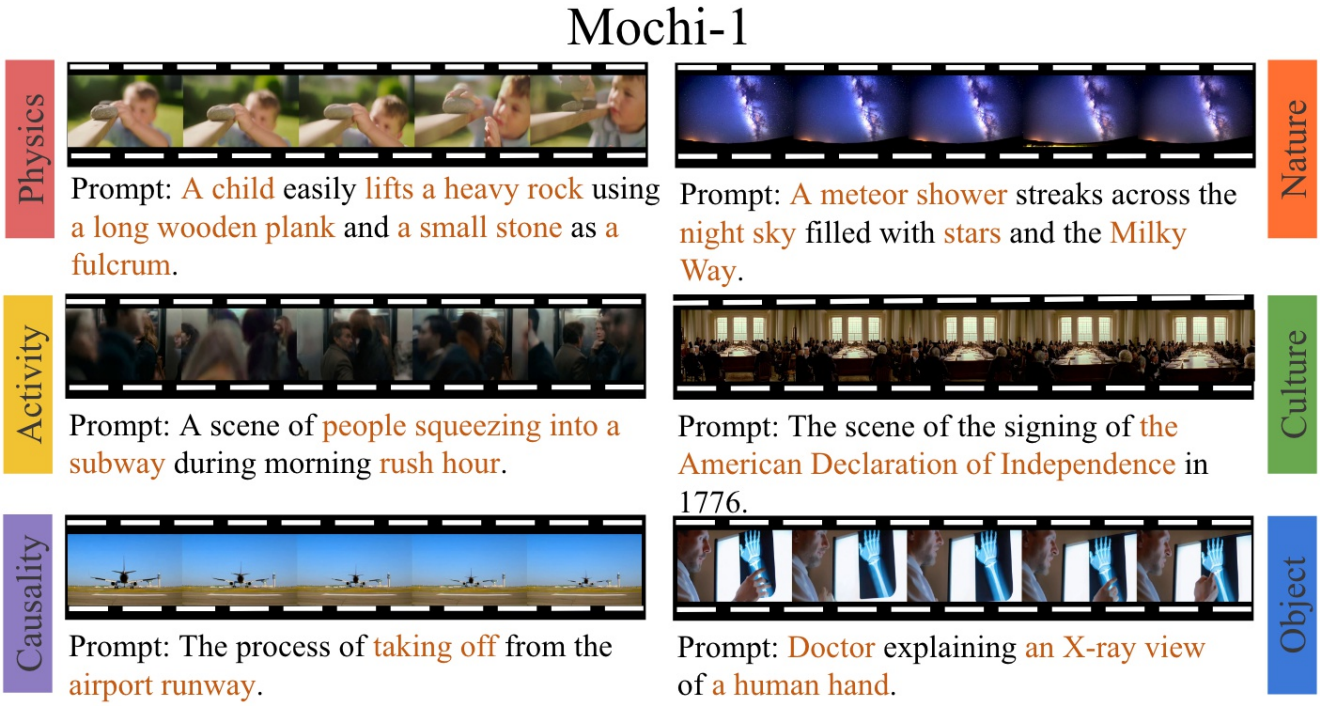


Figure 14. Video generation of Mochi-1.

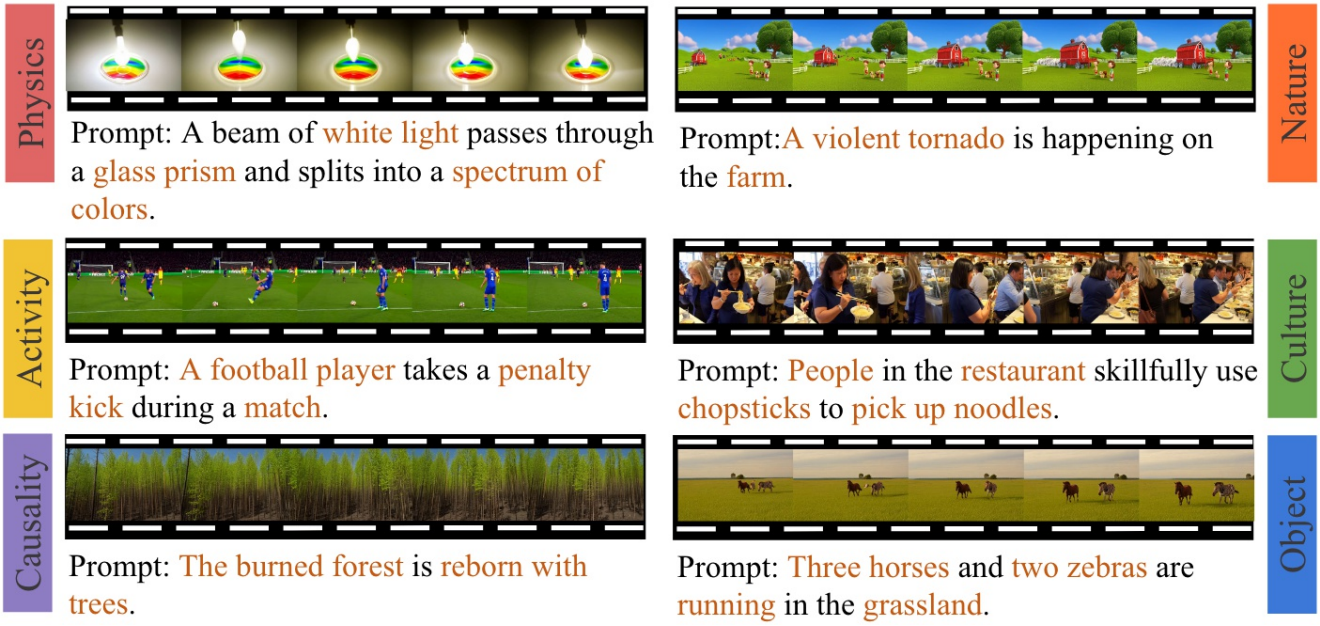


Figure 15. Video generation of Hailuo.


Figure 16. Video generation of Dreamina.

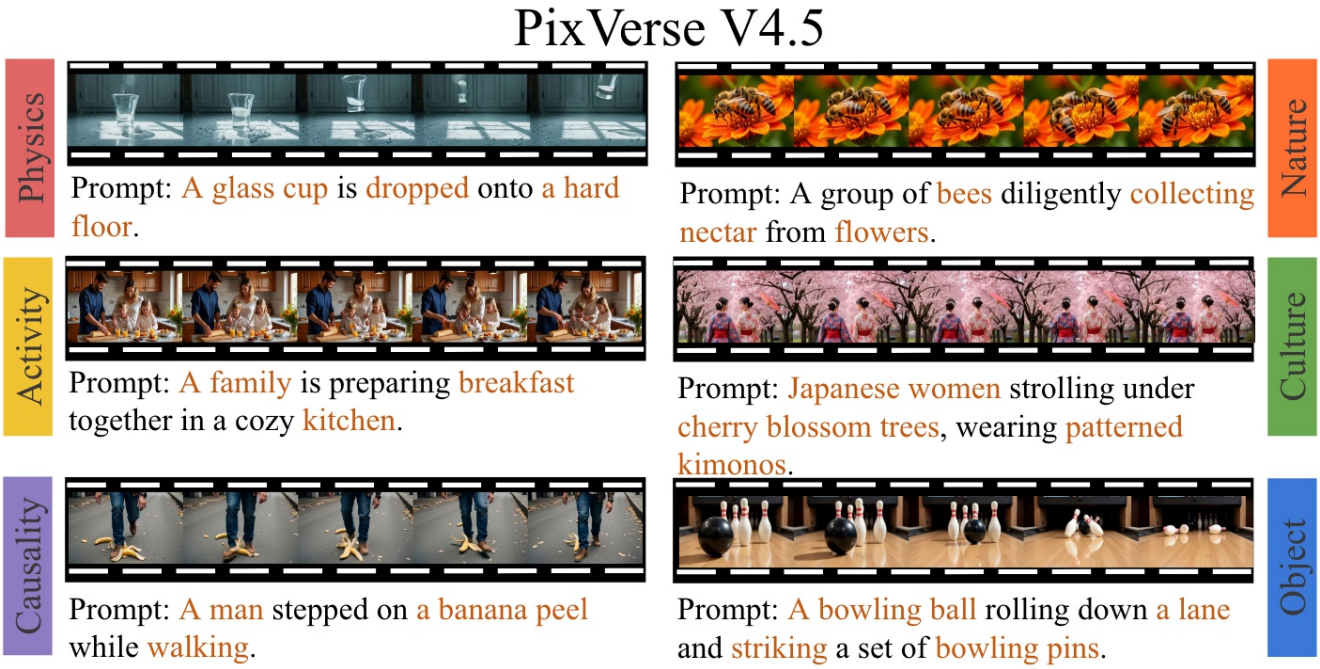


Figure 17. Video generation of PixVerse V4.5.

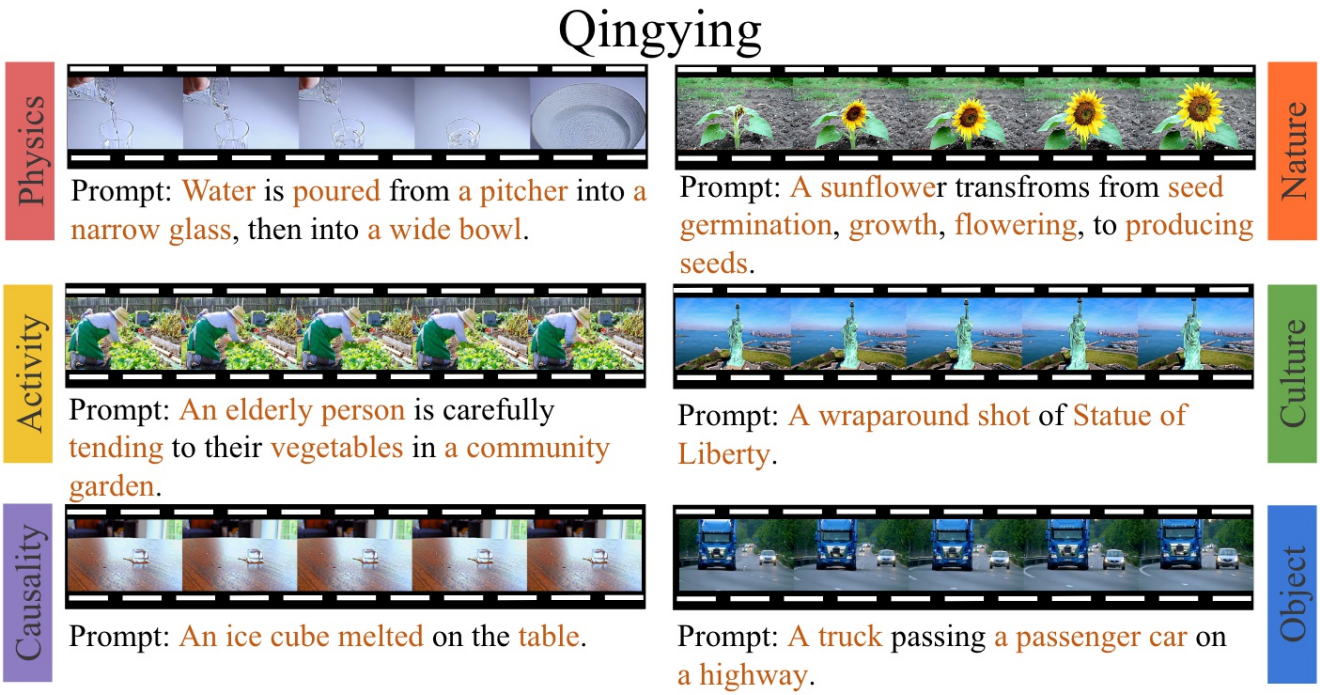


Figure 18. Video generation of Qingying.

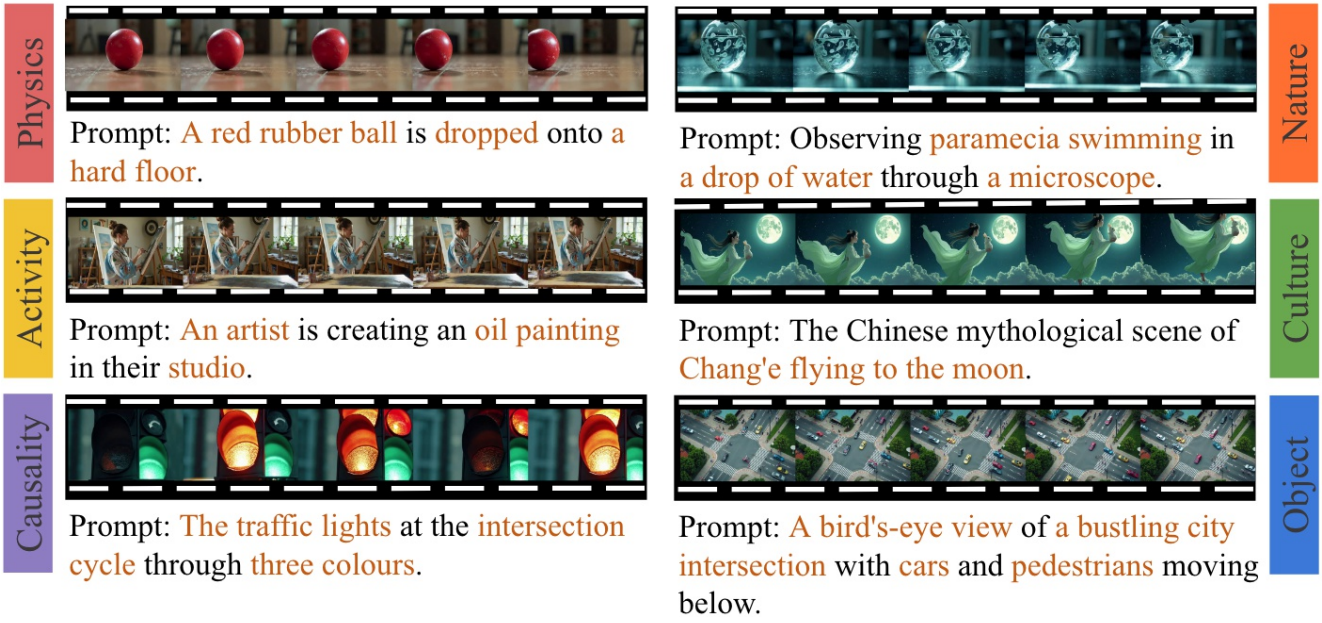


Figure 19. Video generation of LTX Video.

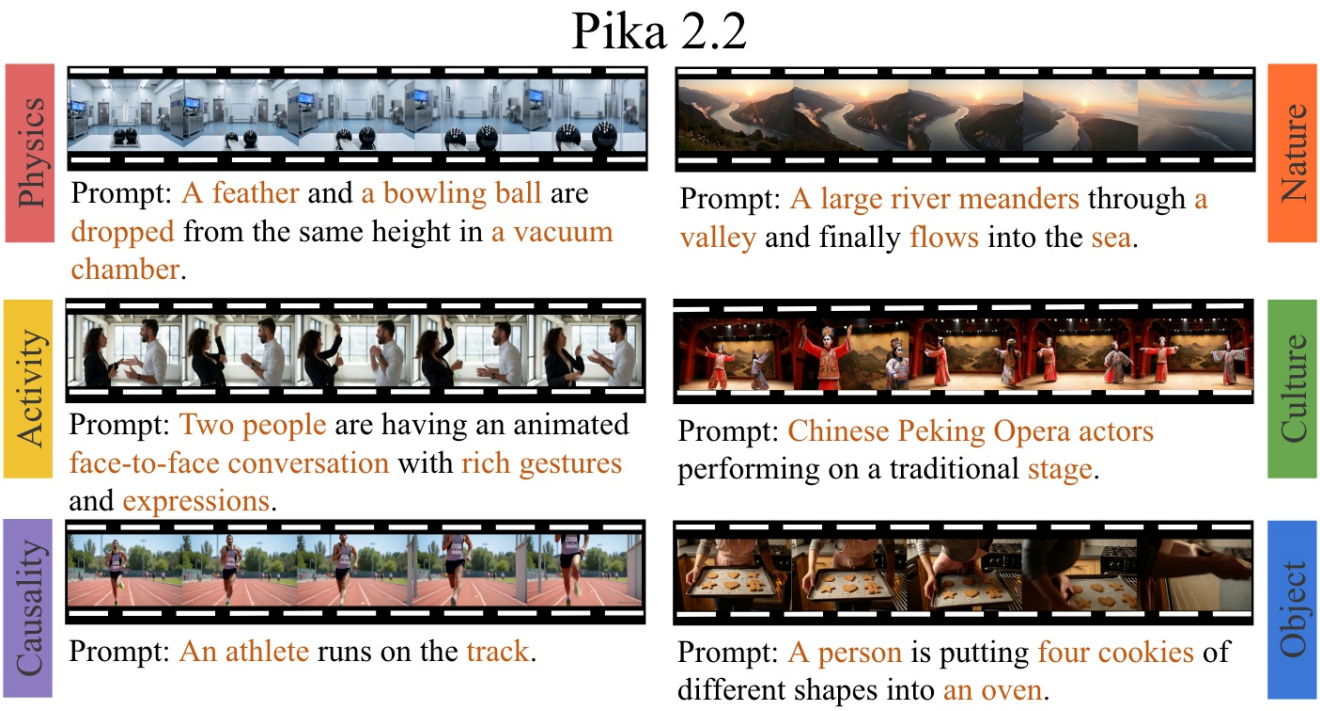


Figure 20. Video generation of Pika 2.2.