Table 7. **Robustness of V-JEPA Surprise Design.** Comparison showing VJEPA architecture and reward hyperparameter against `WMReward(BoN)` results. We report results using 16 particles.

Arch Size	Window	Context	Stride	FPS	Final Score↑
ViT-giant	32	16	16	24	**60.78**
ViT-giant	16	8	8	16	60.34
ViT-giant	16	8	8	24	60.09
ViT-giant	32	16	8	24	60.05
ViT-huge	48	24	8	24	60.04
ViT-giant	16	8	4	24	59.91
ViT-huge	16	8	8	24	59.77
ViT-huge	16	8	4	24	59.62
ViT-huge	32	16	16	24	57.84
ViT-huge	32	16	8	24	57.09

A. Further Ablations

How robust is the `WMReward` to VJEPA size and hyperparameters? One key design factor in our experiments is the choice of hyperparameters for the VJEPA surprise reward $r(\cdot)$ in Equation (6). We study how robust this reward is when transferring physics understanding to video generation. As shown in Table 7, such transfer remains relatively stable across the context length C, prediction horizon M, and stride s. We also observe that physics plausibility gains scale with the size of the reward model (from `ViT-huge` to `ViT-giant`), suggesting that stronger VJEPA backbones yield better performance without any fine-tuning of the underlying video generator.

B. Implementation Details

In the following, we touch on the implementation details of `WMReward`, including generation settings and adaptation to different generation paradigms.

B.1. Generation Settings

For all experiments, we use a vLDM transformer with a spatiotemporal VAE for compression and text–video alignment, and MAGI-1-24B [52], an autoregressive diffusion video model that generates videos chunk-by-chunk using block-causal attention for long-horizon consistency. Their corresponding generation hyperparameters are as follows.

Table 8. **Generation hyperparameters.**

	VideoPhy		*PhysicsIQ*		
Hyperparameter	MAGI-1	vLDM	MAGI-1 (I2V)	MAGI-1 (V2V)	vLDM
Height	480	480	720	720	480
Width	832	720	1280	1280	720
Number of frames	48	49	120	120	49
FPS	24	8	24	24	8
Number of steps	16	50	32	32	50
CFG scale	7.5	6.0	7.5	7.5	6.0
Context guidance scale	1.5	–	1.5	1.5	–
guidance frequency	3	3	5	3	1
VJEPA guidance scale	0.005	0.003	0.005	0.005	0.001

For generation resolution, FPS, we use the recommended settings from the official video generative model repository. Also, for CFG and other guidance implemented in MAGI-1, we follow the default settings in official codebase. The number of generated frames varies according to the specification of the evaluation dataset. For PhysicsIQ, the generated video is required to be exactly 5 seconds. Thus, we generate 49 frames with vLDM and trim them to 40 frames under 8 FPS. For MAGI-1, we generate 120 frames under 24 FPS. For VideoPhy, while there is no explicit requirement on duration and number of frames, we follow the paradigm in the official code to generate relative short video clips with the specification shown. The guidance frequency indicates the time-step interval in which we apply guidance during the denoising process. For vLDM, we use a DDIM [54] scheduler. For MAGI-1, we use standard linear rectified flow sampler. For MAGI-1, we use the distilled 24B checkpoint, and run the inference on 8 H200 GPUs in parallel. For vLDM, we run the inference on a single H200 GPU. Also, for sampling with `WMReward`, we use a VJEPA2 ViT-giant model. The input frame size is 256×256. We choose window size, context length, and stride to be 16, 8, 8 for all experiments.

B.2. Adaptation to Different Generation Paradigms

Current video diffusion models follow two predominant paradigms: holistic generation [1, 59], which denoises all frames simultaneously at the same noise level, and autoregressive generation [52], which generates videos sequentially in temporal chunks, each with its own noise schedule. We implement `WMReward` on models based on both paradigms (vLDM and MAGI-1). Practically, in both cases, the BoN search implementation is the same, while the implementation of guidance (∇) varies. For holistic generation, we use the the combined CFG and `WMReward` guidance in Equation (13).

$$\begin{aligned}\nabla_{x_t} \log p_\lambda\big(x_t \mid \text{txt}\big) = (1-\omega_{\text{txt}})\,\nabla_{x_t} \log p(x_t) \\ + \omega_{\text{txt}}\,\nabla_{x_t} \log p\big(x_t \mid \text{txt}\big) \\ - \omega_s \nabla_{x_t} r\big(x_t \mid \text{txt}\big).\end{aligned} \tag{13}$$

Specifically, we adopt a sliding window approach to split the video into context and prediction target chunks, compute the VJEPA surprise on each window, and average over the whole sequence. For autoregressive generation, we perform guidance as follows:

$$\begin{aligned}\nabla_{x_t} \log p_\lambda\big(x_t \mid x_t^{<k}, \text{txt}\big) = (1-\omega_{<k})\nabla_{x_t} \log p(x_t) \\ + (\omega_{<k}-\omega_{\text{txt}})\nabla_{x_t} \log p\big(x_t \mid x_t^{<k}\big) \\ + \omega_{\text{txt}}\nabla_{x_t} \log p\big(x_t \mid x_t^{<k}, \text{txt}\big) \\ - \omega_s \nabla_{x_t}\, r\big(x_t \mid x_t^{<k}, \text{txt}\big),\end{aligned} \tag{14}$$

where we combine VJEPA surprise guidance with classifier-free guidance from both text and previous denoised chunks $x_t^{<k}$. In particular, we use previous denoised

chunks as context for the VJEPA predictor, predict the next chunk, and calculate VJEPA's surprise reward.

B.3. VLM-based Reward Model Details

For VLM-based reward models, we use Qwen2.5-VL-7B-Instruct [5] and Qwen3-VL-8B-Instruct [65], respectively. We use a question template `"Does the video show good physics dynamics and showcase a good alignment with the physical world? Please be a strict judge. If it breaks the laws of physics, please answer 0. Answer 0 for No or 1 for Yes. Reply only 0 or 1."`. Then, we extract the logit of token `"1"` and its variation `" 1"` as the reward signal.

C. Human Study Details

As shown in Figure 5, annotators are presented with a side-by-side comparison interface where they view two generated videos along with the original text prompt and conditioning frames describing the physical scenario. For each video pair, annotators provide judgments across three criteria: **(1) Physics Plausibility**, assessing whether the physical interactions and dynamics are realistic; **(2) Visual Quality**, evaluating the overall visual fidelity, clarity, and aesthetics of the generated video; and **(3) Prompt Alignment**, measuring how well the video content matches the given text description. For each criterion, the annotators select one of three options: preferring the video sample on the left, preferring the video sample on the right, or reporting a neutral preference when the difference is negligible or both videos are equally good/bad. To mitigate position bias, videos from each model are randomly assigned to the left or right positions. We collect evaluations from five annotators. Results are aggregated to compute win rates (percentage of comparisons where a model is preferred, excluding neutral judgments) and accuracy scores (computed as $\frac{\text{wins}+0.5\times\text{neutrals}}{\text{total}}$), providing a comprehensive assessment of relative model performance.

D. Qualitative Examples

We present additional qualitative visual samples to demonstrate the effectiveness of `WMReward`. As shown in Figures 6 to 8 for image- and multiframe-conditioned generation, and in Figure 9 for text-conditioned generation, we observe that our generated videos exhibit improved physics plausibility across spatial continuity, rigid-body dynamics, fluid behavior, buoyancy, temporal continuity, gravity, conservation of mass, and optical effects. These samples indicate that the latent world model contains non-trivial physics understanding that enables more physically plausible video generation. *For a better view of the dynamics, we recommend viewing the videos attached in the supplementary material zip file.*

E. Failure Mode Analysis

One of our core assumptions is that the latent world model VJEPA-2 captures stronger intuitive physics than current video generators [23]. However, this learned prior remains a proxy for true physics dynamics. VJEPA surprise is not exclusively measuring physics plausibility and entangles other perceptual factors. Practically, as shown in Figure 10, we observe that sampling with VJEPA reward in some cases does not lead to substantial physics plausibility improvements. For example, the model fails to capture abrupt physical events, such as fluid overflowing from a bottle or a lit match igniting a balloon and causing it to explode, which require reasoning about sudden state changes. It sometimes struggles with more complex phenomena, including mirror reflections and siphon effects, which demand more complex reasoning and understanding of material properties. While VJEPA reward can correct some physics violations (*e.g.*, conservation-of-mass errors in certain siphon scenarios), these failures indicate that there remains room to improve the physics understanding of latent world models in order to obtain more reliable reward signals and, consequently, better physics-aware video generation.

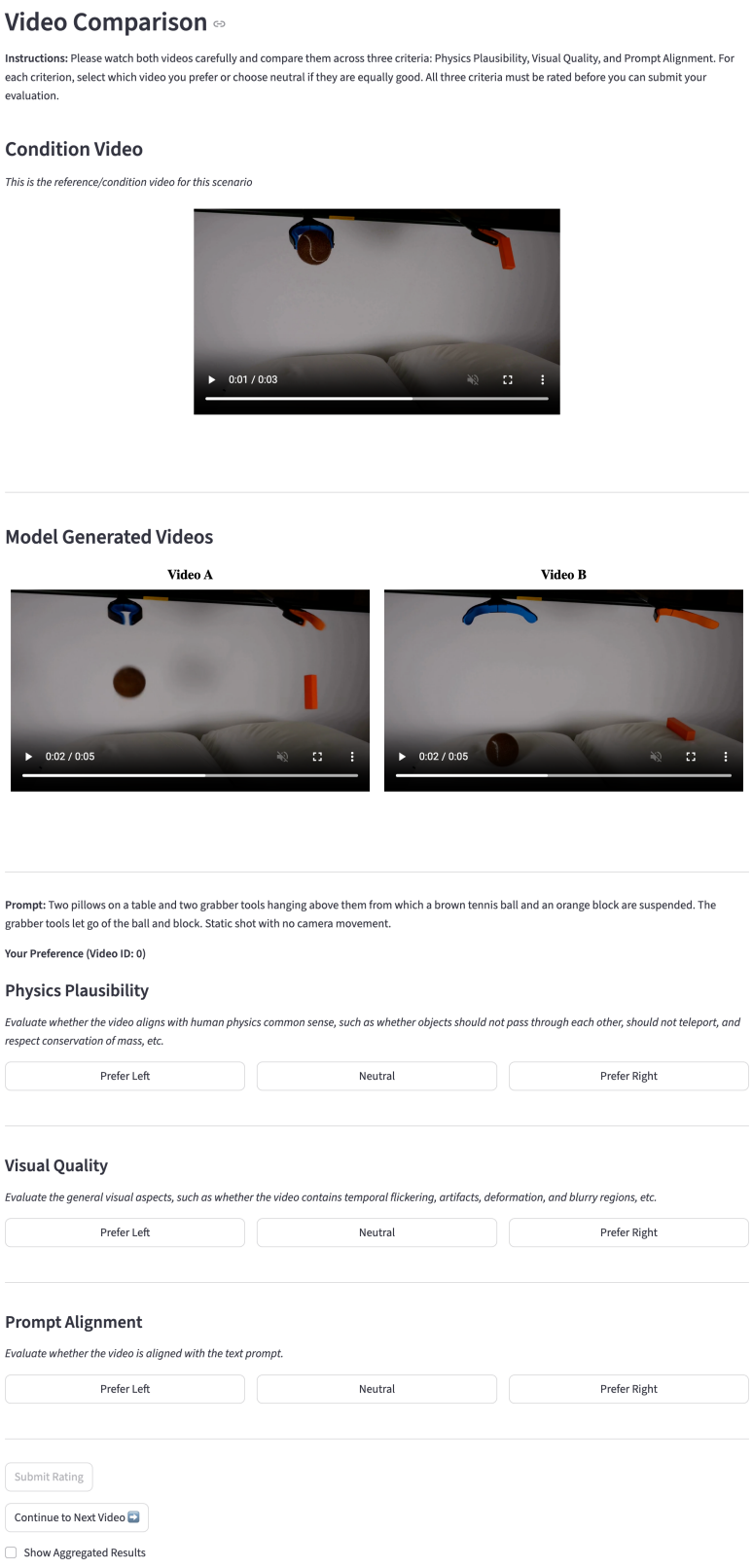

Figure 5. **Human Study Interface.** Annotators view a side-by-side video comparison and indicate their preference on three criteria—Physics Plausibility, Visual Quality, and Prompt Alignment—choosing one of three preference options: Left, Right, or Neutral.

Figure 6. **Additional Qualitative Results on Physics-IQ.**

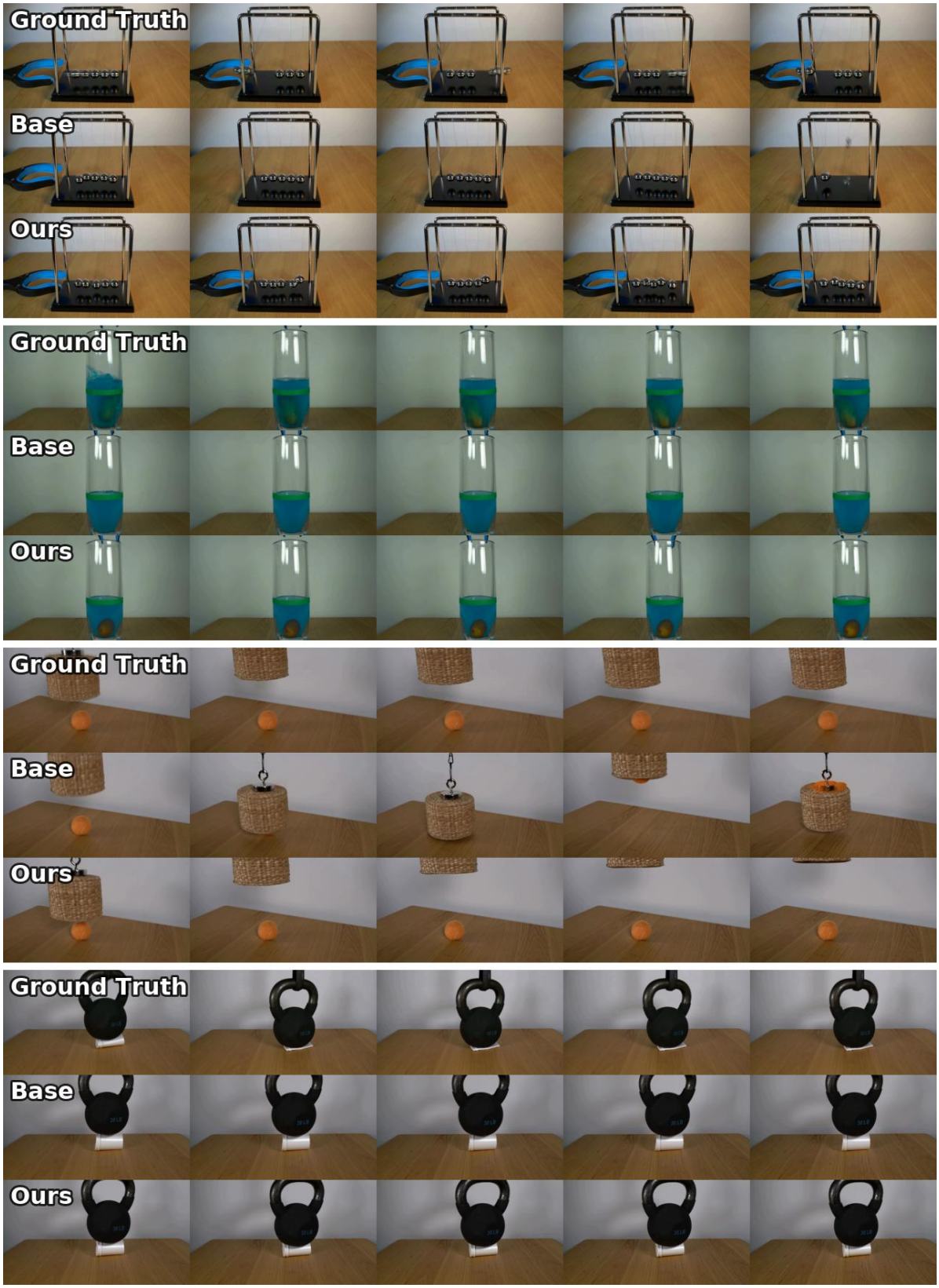

Figure 7. **Additional Qualitative Samples on Physics-IQ.**

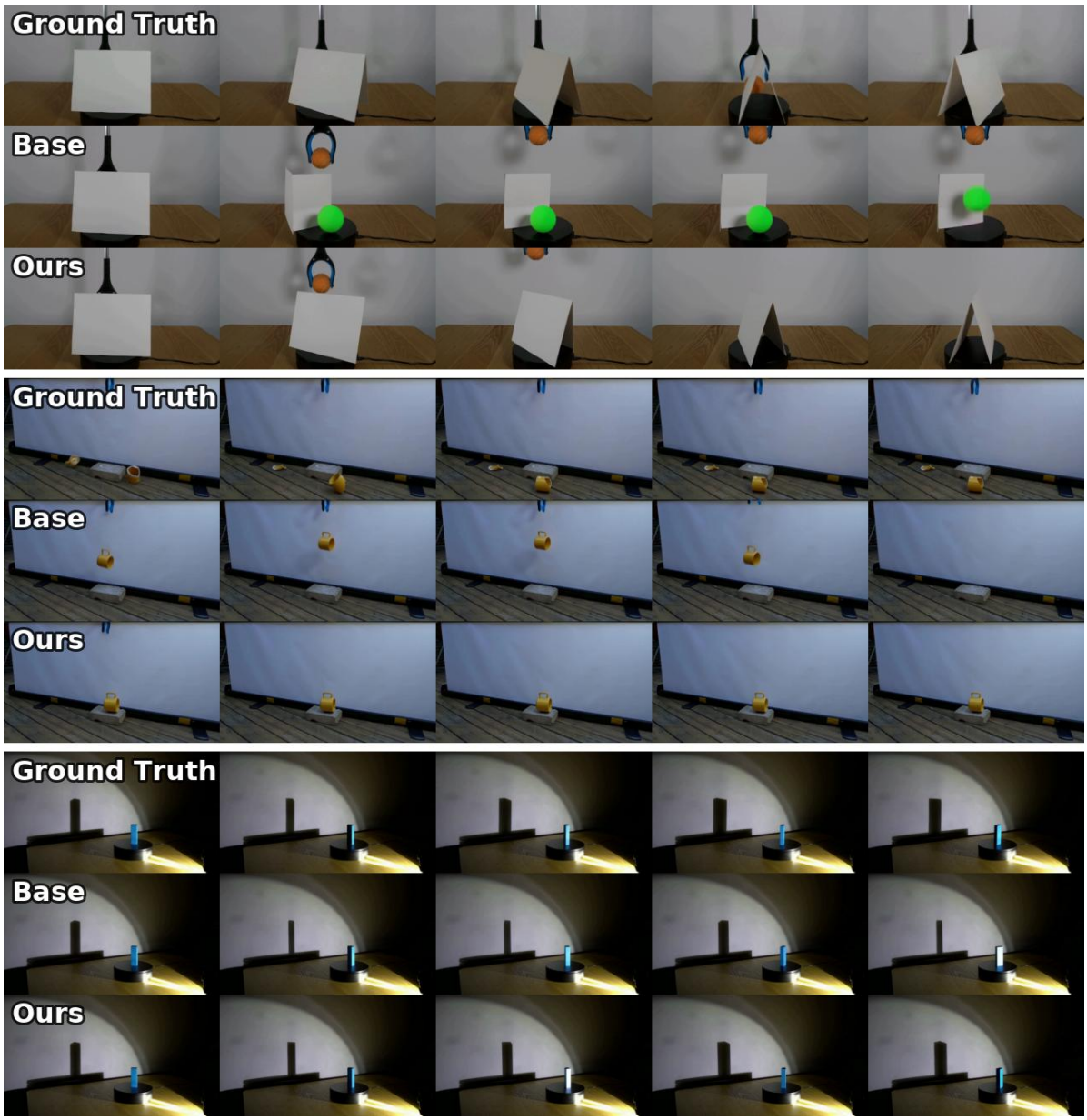

Figure 8. **Additional Qualitative Samples on Physics-IQ.**

Figure 9. **Additional Qualitative Samples on VideoPhy.**

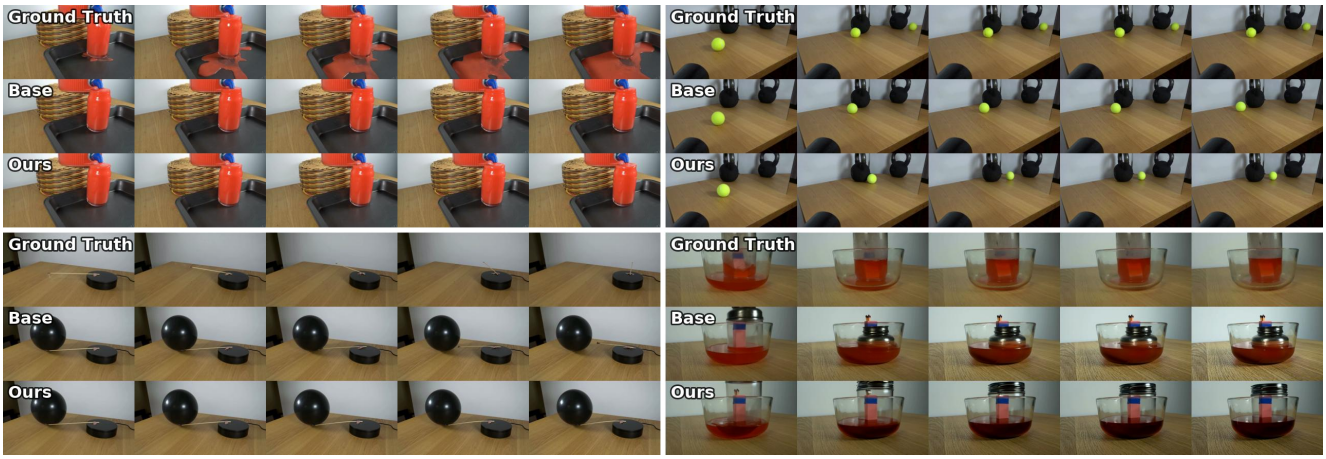

Figure 10. **Failure Mode Analysis.** We observe some failure modes that persist even when leveraging VJEPA-2 for sampling. For example, the model often fails on abrupt physical events, such as fluid overflowing from a bottle (top left quadrant) or a lit match igniting a balloon and causing it to explode (bottom left quadrant). It also struggles with more complex phenomena that requires reasoning and understanding of material properties, including mirror reflections (top right) and siphon effects (bottom right), indicating that both the base model and the reward model still have room of improvement on physics understanding.