VIPER: VISUAL IN-CONTEXT PHYSICS REASONING FOR PHYSICALLY PLAUSIBLE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

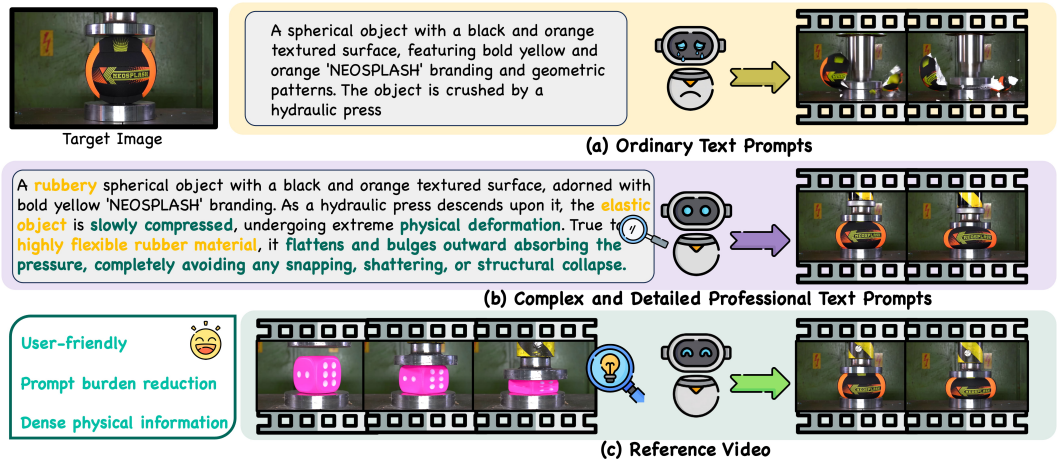


Figure 1: **Motivation**. (a) With an ordinary prompt, a video generator may produce visually plausible but physically implausible results because the physical process is under-specified. (b) A detailed and carefully engineered prompt can better communicate the expected physical behavior, but such prompts are difficult to write in practical use. (c) VIPER uses a reference video as a dense visual in-context demonstration, allowing the model to generate the target video with the expected physical behavior without requiring an exhaustive textual description.

ABSTRACT

Modern video generation models can synthesize visually compelling and temporally coherent clips, yet controlling their physical behavior remains difficult with standard text and image conditions. The core challenge is a **conditioning bottleneck**: material response, contact interaction, deformation, and motion trajectory are continuous and relational physical cues that are hard to specify exhaustively in language but can be demonstrated naturally by video. We propose **VIPER**, a Visual In-Context Physics Reasoning framework for reference-guided image-to-video generation. Given a target image, a brief target prompt, and a reference video, VIPER treats the reference as a dense visual demonstration of the desired physical process rather than an appearance template. It uses a Multimodal Large Language Model (MLLM) to extract reference-derived physical cues and guide a pretrained image-to-video generator through a hierarchical training strategy, enabling physical behavior transfer while preserving the visual prior of the base generator. To support this setting, we construct **VIPER-19K**, a curated dataset with material, trajectory, and physical-impact annotations, together with filtered reference-target pairs. Experiments on an unseen validation set show that VIPER achieves stronger reference-video physical similarity and higher human preference than representative video generation and video-as-prompt baselines, while maintaining competitive general video quality. Qualitative results further demonstrate that VIPER can transfer reference-derived physical behavior to new target scenes without requiring carefully engineered prompts.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Recent video generation models (Brooks et al., 2024; Kong et al., 2024; Wan et al., 2025) have rapidly improved in generating visually plausible and temporally coherent videos, benefiting from large-scale training data, increased model capacity, and advances in diffusion-transformer architectures (Peebles & Xie, 2023). Despite substantial progress in visual fidelity and temporal coherence, generating videos that follow real-world physical regularities remains challenging, as physical behavior is difficult to communicate through ordinary prompts and images alone. The community has explored several directions toward physically plausible generation, including simulator-assisted generation (Foo et al., 2026; Liu et al., 2026; Li et al., 2025), physics-controllable synthesis (Wang et al., 2025a), and learning physical priors from physics-aware data (Wang et al., 2025b; Gillman et al., 2025).

However, we observe that the difficulty in this setting lies less in video synthesis itself than in understanding and specifying the physical behavior to be generated. Specifically, conventional image-to-video (I2V) models may generate physically implausible results from ordinary prompts, while sometimes producing realistic physical processes when material properties, motion trajectories, and physical impacts are explicitly specified, as illustrated in Fig. 1(a) and (b), respectively. However, such physically detailed prompts, which must specify continuous and relational cues such as where deformation starts, how contact alters motion, how materials respond over time, are difficult to write. This observation motivates our view that generating physically plausible video faces a **conditioning bottleneck**: video generation models may already contain useful physical priors learned from large-scale data, but existing text and image conditions provide only limited capacity for reliably specifying complex physical processes. This bottleneck suggests a more natural alternative: physical processes are often easier to observe visually than to describe exhaustively, since their dynamics can be directly demonstrated and perceived in video. A more natural interface for physical control is **reference-guided I2V generation**. Instead of enumerating every physical detail in text, a reference video can serve as a dense visual demonstration of how a physical process unfolds, as shown in Fig. 1(c). The target image and a brief prompt define *what* should appear in the desired scene, while the reference video demonstrates *how* the physical behavior should evolve. The resulting task turns physically plausible video generation into a form of **visual in-context physics reasoning**: the model must infer material response, motion trajectory, and object interaction from a visual example and apply them to a new target scenario.

To instantiate this motivation, we propose **VIPER**, a Visual In-Context Physics Reasoning framework for physically plausible video generation. Given a target image, a brief prompt, and a reference video, VIPER does not aim to copy the visual appearance of the reference video, but instead transfers its implicit physical pattern to the target image. This design introduces a physically informative condition through visual observation while keeping the text-conditioning interface simple and intuitive. To effectively extract the implicit physical information embedded in the reference video, VIPER leverages the visual understanding capability of a multimodal large language model (MLLM) to interpret reference-derived physical cues, which are then used to guide an I2V generator in producing the corresponding target physical behavior. Moreover, using a reference video as a physical condition requires the model to identify which aspects of the reference are transferable. Specifically, the model must distinguish physical behavior from irrelevant appearance such as object category and scene layout. VIPER therefore adopts a hierarchical training strategy that progressively learns to interpret physical cues, transfer them across target scenes, and preserve the visual quality of the pretrained generator. This staged design encourages the model to use the reference video as a source of physical behavior rather than as a pixel-level template.

Furthermore, existing video datasets (Lin et al., 2024; Wang & Yang, 2024) are not well suited to this setting, as they are primarily designed to capture semantic diversity or general visual quality, rather than to provide structured supervision over physical factors such as material properties, motion trajectories, and physical impacts. To fill this gap, we construct **VIPER-19K**, a primarily real-world curated dataset for reference-guided physical video generation. Unlike existing datasets (Lin et al., 2024; Wang & Yang, 2024), VIPER-19K integrates high-quality internet videos, cleaned public physics-oriented datasets, commercial model generations, and simulator-based samples. Each video is annotated along material, trajectory, and physical-impact dimensions, and reference-target pairs are constructed by grouping samples with compatible physical attributes and further filtering coarse

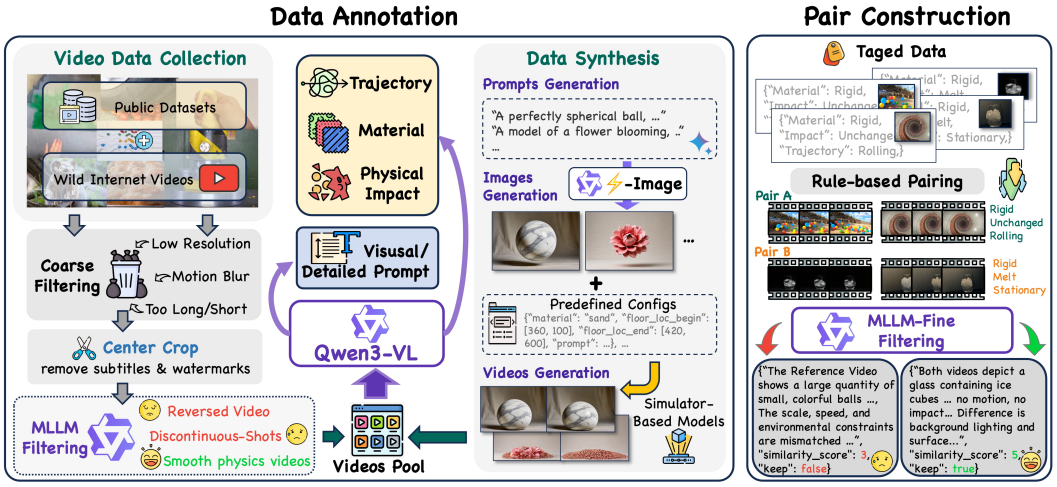


Figure 2: Overview of the VIPER-19K construction pipeline. (a) We collect candidate videos from internet sources, public datasets, and physics-based simulation pipelines, then remove noisy clips and annotate each retained video with material, trajectory, and physical-impact attributes. (b) We construct reference-target video pairs by first grouping videos with compatible physical attributes and then filtering with a MLLM to retain pairs with transferable physical behavior.

pairs using a MLLM. The resulting dataset supports learning transferable physical behavior from reference videos, rather than relying on object-category or appearance shortcuts.

Experiments validate the effectiveness of reference-guided physical conditioning across quantitative metrics and qualitative comparisons. VIPER more reliably transfers reference-derived physical behavior to new target scenes while maintaining competitive visual quality. In summary, our contributions are:

- We formulate **visual in-context physics reasoning** for reference-guided I2V generation, where a reference video provides dense and transferable physical cues that are difficult to specify completely with text prompts.
- We propose **VIPER**, a reference-guided generation framework that uses visual understanding to extract transferable physical cues from a reference video and guides target video synthesis through a hierarchical strategy.
- We construct **VIPER-19K**, a curated physical video dataset with material, trajectory, and physical-impact annotations as well as filtered reference-target pairs for learning visual physical transfer.

2 Related Work

2.1 Video Generation and Physical Plausibility

Video generation has progressed from adapting image diffusion priors to developing large-scale video-native generative models. Early methods (Guo et al., 2024; Chen et al., 2024) extend text-to-image diffusion models with temporal modeling capabilities through motion adaptation and improved video diffusion training. Recent architectural and training advances, including diffusion transformers (Peebles & Xie, 2023), rectified-flow formulations (Esser et al., 2024), and flow-matching objectives (Lipman et al., 2022), further scale training data, model capacity, and training recipes, leading to increasingly capable text-to-video models (Yang et al., 2024; Kong et al., 2024; Wan et al., 2025). While this scaling trend has substantially improved visual fidelity and spatiotemporal consistency, it does not effectively ensure physically plausible material responses, motion trajectories, or object interactions, leaving physical plausibility a persistent challenge for current video generation models.

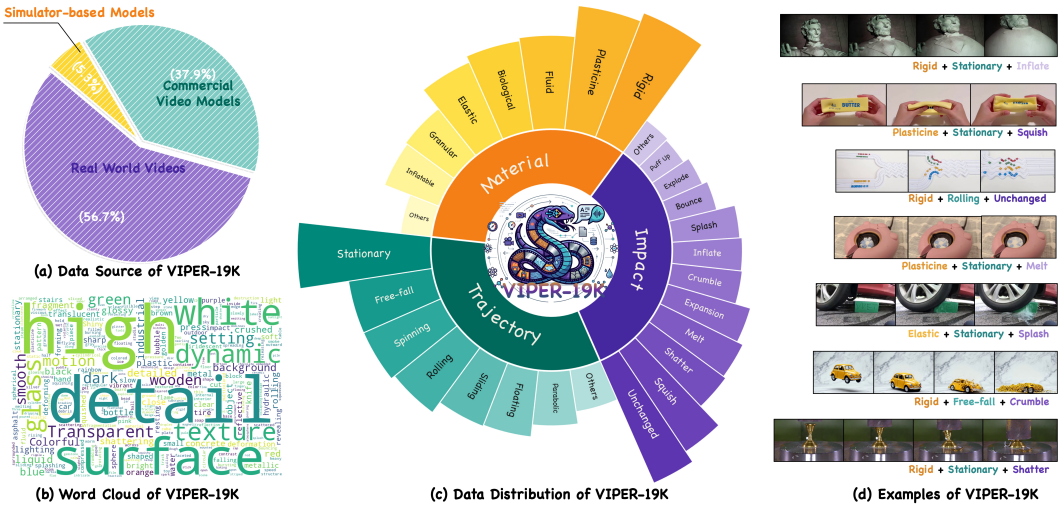


Figure 3: Statistics and examples of VIPER-19K. **(a) Data sources** across real-world videos, commercial video-model generations, and simulator-based samples; **(b) Prompt word cloud** of the collected data; **(c) Attribute distribution** coverage over materia, trajectory, and physical-impact labels and **(d) Diverse video examples** with material–trajectory–impact triplets.

Prior work improves physical behavior by introducing simulation, control signals, or physics-aware data. Simulator-based and physics-controllable systems (Jiang et al., 2016; Macklin et al., 2016; Authors, 2024; Foo et al., 2026; Liu et al., 2026; Li et al., 2025; Zhang et al., 2025a) couple generators with physics engines, simulation-inspired constraints, part-aware grounding, or action-conditioned signals. Other methodss (Wang et al., 2025a; Gillman et al., 2025; 2026; Yuan et al., 2025; Wang et al., 2025b; Shen et al., 2026) learn physical priors from simulated trajectories, synthetic examples, neural dynamics, real-world physical videos, or latent physical representation. These approaches show the value of physical supervision, but often rely on simulator access, predefined control formats, object-specific physical states, or task-specific annotations. VIPER instead treats a reference video as the physical specification, allowing the model to infer transferable physical attributes from visual evidence without requiring explicit forces, trajectories, or simulator states as input.

2.2 In-Context Learning for Video Generation

In-context conditioning offers a flexible alternative to manually specifying generation behavior. One group of methods uses visual examples for instructional editing, reference-guided creation, or video editing (Zhang et al., 2025b; Jiang et al., 2025; Ye et al., 2025; Wei et al., 2025). Another group adapts diffusion transformers or control interfaces to accept in-context conditions with minimal task-specific modification (Huang et al., 2024a; Tan et al., 2025), while Video-as-Prompt (Bian et al., 2025) treats videos as prompts for unified semantic control. These methods show that visual examples can provide rich conditioning signals beyond text, but their primary focus is semantic layout, appearance, editing intent, or general motion consistency. VIPER differs by treating the reference video as evidence of physical behavior: the example is used to specify transferable material response, motion trajectory, and interaction dynamics for target I2V generation.

3 VIPER-19K Dataset

Visual in-context physics reasoning requires not only videos with observable physical phenomena but also **reference-target pairs** whose behavior can be transferred across different appearances and scenes. Existing video resources are poorly matched to this requirement. Web videos often contain watermarks, abrupt edits, reversed clips, or ambiguous dynamics, while general generation datasets prioritize semantic coverage over material response, motion trajectories, and physical interactions. We therefore construct **VIPER-19K**, a curated dataset for training and evaluating reference-guided physical video generation.

3.1 Data Collection and Annotation

VIPER-19K is curated from three complementary sources, encompassing real-world videos, synthetic videos generated by commercial video generation models and simulation-based models. First, we collect in-the-wild web videos that capture explicit physical events, including compression, deformation, collision, melting, and crumbling. Second, we curate and integrate data from WISA-80K (Wang et al., 2025b) and VAP (Bian et al., 2025); the former offers diverse real-world footage, whereas the latter supplies out-of-distribution synthetic videos generated by commercial models. Third, we augment our dataset by synthesizing additional samples using physics-based pipelines, such as PhysCtrl (Wang et al., 2025a) and RealWonder (Liu et al., 2026), to ensure explicitly controlled physical interactions.

We then filter out clips with severe compression, watermarks, shot transitions, reversed temporal order, or ambiguous physical content. Each retained video is annotated along three axes: ***material***, ***trajectory***, and ***physical impact***. These labels describe object response properties, dominant motion patterns, and physical events, and are used for both dataset analysis and pair construction.

3.2 Pair Construction

The core supervision unit of VIPER-19K is a reference-target video pair. The reference video demonstrates physical behavior, while the target video provides generation supervision under a different visual condition. Coarse labels are useful for retrieval but insufficient for pairing, because two clips with the same event type may still differ in material response, motion direction, or interaction strength.

We construct pairs in two stages. First, we group videos by compatible material, trajectory, and physical-impact labels, then sample candidate pairs within each bucket. Second, we use Qwen3-VL-32B to assess whether the physical behavior in the reference video is transferable to the target video. Pairs with only superficial semantic overlap or inconsistent physical dynamics are removed, yielding cleaner supervision for visual in-context learning.

3.3 Data Diversity

Fig. 3 summarizes the source and attribute coverage of VIPER-19K. The dataset combines real-world observations, commercial-model generations, and simulation-based samples, exposing the model to natural visual complexity, synthetic out-of-distribution cases, and controlled physical variations. Its annotations span multiple materials, trajectories, and physical-impact categories, making pair sampling less tied to object category or appearance and encouraging the model to learn transferable physical behavior.

4 Method

4.1 Overview

Physically plausible image-to-video (I2V) generation requires control over both visual content and dynamic behavior. Standard I2V inputs provide only a sparse interface for this control: the target image fixes appearance, while an ordinary prompt often leaves material response, contact evolution, and motion details underdetermined. To provide richer physical guidance, we define **reference-guided I2V generation**, where the target image I^{tar} and a simple target prompt T^{tar} specify the desired content, and a reference video V^{ref} provides the physical behavior to transfer. The desired output V^{tar} should preserve the appearance and semantics of the target condition while following physical attributes observed in the reference video.

VIPER augments the target image and prompt with a visual interface for physical control. As shown in Fig. 4, VIPER first uses a multimodal large language model (MLLM) to distill physics-relevant cues from V^{ref} into compact conditioning tokens. These tokens are then injected into a DiT-based I2V diffusion model together with the target prompt, while the target image is provided through the original I2V pathway. In this way, the target image and prompt define what should appear in the output, and the reference video specifies how the physical behavior should unfold.

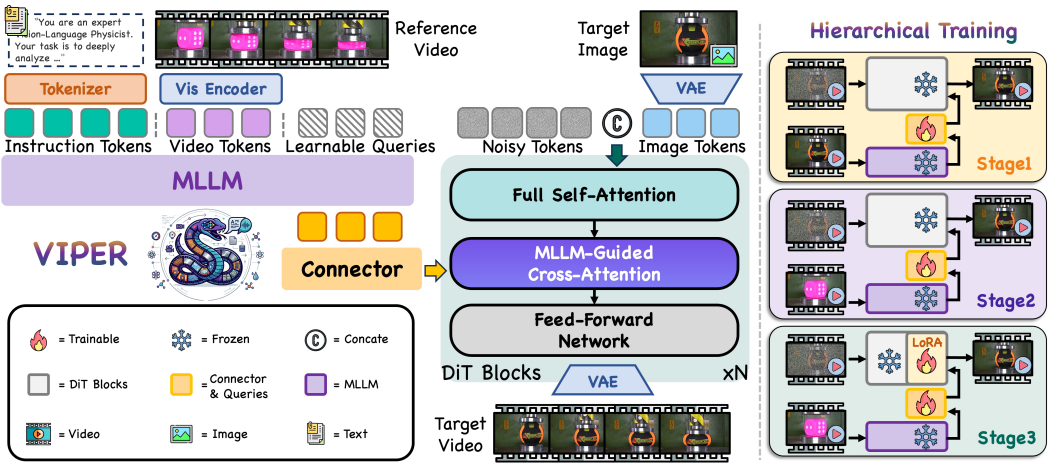


Figure 4: Overview of VIPER. **(a) The pipeline of VIPER:** The reference video is encoded by an MLLM with learnable physics query tokens $q_{\text{learnable}}$, the corresponding hidden states v are projected into condition tokens c, and the DiT-based I2V generator synthesizes the target video from the target image, prompt, and reference-derived physical conditions. **(b) The hierarchical training strategy:** VIPER first aligns the MLLM-to-DiT conditioning interface, then learns reference-target physical transfer, and finally adapts the generator with lightweight trainable parameters.

4.2 Reference-Video Physical Conditioning

The key design problem is how to convert a reference video into a conditioning signal that a DiT-based I2V diffusion model can use. Physical cues in V^{ref} are distributed across frames and involve coupled changes in object state, motion, and interaction. Reducing them to hand-defined labels would discard fine-grained dynamics. VIPER therefore uses the MLLM as a **visual physics encoder** and introduces learnable physics query tokens $q_{\text{learnable}}$ to extract a compact physical context from V^{ref}.

Given V^{ref}, we first convert the video into visual tokens with the MLLM visual encoder. These visual tokens are concatenated with system prompt tokens p_{sys} and learnable query tokens $q_{\text{learnable}}$. The system prompt biases the MLLM toward physical behavior rather than object identity or scene appearance, while $q_{\text{learnable}}$ serves as trainable slots for gathering generation-relevant dynamics. After the MLLM forward pass, we take the hidden states of $q_{\text{learnable}}$ as the reference physics representation v^{ref}:

$$v^{ref} = \text{MLLM}(V^{ref}, p_{\text{sys}}, q_{\text{learnable}}), \tag{1}$$

This query-based representation avoids passing all MLLM hidden states to the diffusion model directly, which would mix physical information with task-irrelevant visual and textual tokens and add unnecessary computational overhead.

To make the MLLM representation compatible with the DiT model, we use a connector g_θ to project v^{ref} into physics condition tokens c_p^{ref}:

$$c_p^{ref} = g_\theta(v^{ref}). \tag{2}$$

Building on the original DiT input structure (Wan et al., 2025), we compress the target image I^{tar} via a VAE (Kingma & Welling, 2013) for concatenation with the noisy latent tokens. Concurrently, the target prompt T^{tar} is encoded by a T5 encoder (Raffel et al., 2020) and concatenated with the reference-derived physics condition tokens.

4.3 Hierarchical Training

Directly training VIPER on reference-target pairs is unstable before the MLLM representation space is aligned with the DiT latent space. Such instability may causes the diffusion model to overfit to the input target image rather than inferring physical dynamics from the reference information. We



Figure 5: Qualitative comparison on representative examples. Each case provides a reference video that demonstrates the desired physical behavior and a target image that defines the generated content. In the melting case, it captures both the geometric change and the gradual color transition of the melted fluid; In the deformation case, it follows the reference motion pattern and generates a can that folds and twists from the physically plausible region.

therefore adopt a three-stage **hierarchical training** strategy that decouples cross-model alignment, physical transfer learning, and lightweight generator adaptation.

The first two stages train only the learnable query tokens and connector while keeping both the MLLM and DiT frozen. In **Stage 1**, we utilize the target video itself as the MLLM input. This self-referential setup forces the connector to produce generator-compatible conditioning tokens, establishing a stable alignment before true cross-video transfer is required. To prevent the model from exploiting pixel-level shortcuts, we employ a suite of photometric and spatial augmentations, specifically color adjustment, translation, and horizontal flipping. In **Stage 2**, we switch to VIPER-19K reference-target pairs: the MLLM encodes the reference video, and the generator is supervised by the target video. This stage encourages the query representation to capture physical attributes that can transfer across videos. In **Stage 3**, to enhance the model's sensitivity to the injected physics tokens, trainable LoRA (Hu et al., 2022) parameters are introduced into selected DiT layers. Concurrently, the query tokens and the connector undergo further optimization on reference-target pairs. Overall, the staged design separates representation alignment from generator adaptation and preserves the visual prior of the pretrained video generator.

5 Experiments

5.1 Experimental Setup

Implementation details. We train VIPER on top of Qwen3-VL-4B-Instruct (Bai et al., 2025) and Wan2.2-I2V-14B (Wan et al., 2025). Following the pretrained DiT video generator, all training videos are resized to 480×832 and sampled at 16 FPS with 81 frames. We use AdamW with a learning rate of 1×10^{-5} for all three training stages. Stage 1 is trained for 3K steps using real-world videos. Stage 2 is trained for 6K steps using synthetic videos from closed-source video models, paired videos generated by simulation-based methods, and a subset of high-quality real-world pairs. Stage 3 is trained for 6K steps using all reference-target pairs. All stages are trained on 6 H20 GPUs. During inference, we use 50 denoising steps and set the classifier-free guidance scale to 6.

Table 1: Quantitative comparison on the evaluation dataset. Red and yellow cells indicate the best and second-best results.

Method	VBench (Huang et al., 2024b)			VLM-as-Judge	User Study
	Motion Smoothness ↑	Temporal Flickering ↑	Aesthetic Quality ↑	Physical Similarity ↑	Preference Rank ↓
Wan2.1-I2V-14B	97.86	96.34	50.23	2.67	2 (33.3%)
Wan2.2-TI2V-5B	98.03	96.97	49.20	2.61	
Wan2.2-I2V-14B	98.58	97.10	50.84	2.97	
VAP-CogVideo	98.58	97.44	50.19	2.92	3 (5.9%)
VAP-Wan	97.32	95.88	49.03	2.87	
VACE	98.73	97.81	49.33	2.14	4 (2.0%)
VIPER (Ours)	**98.60**	**97.45**	**51.86**	**3.42**	**1 (58.8%)**

Evaluation dataset. We evaluate on an unseen validation set constructed alongside VIPER-19K. The full validation set contains 500 reference-target pairs, from which we randomly sample 75 pairs for evaluation.

Baselines. We compare VIPER with video generation fundamental models and in-context learning baselines: Wan2.1-I2V-14B, Wan2.2-TI2V-5B, Wan2.2-I2V-14B (Wan et al., 2025), VAP (Bian et al., 2025), and VACE (Jiang et al., 2025). To ensure a fair comparison and validate our advantage of implicit transfer, for methods that do not natively support in-context video conditioning, we use Qwen3-VL-4B-Instruct (Bai et al., 2025) to extract textual physical descriptions from the reference video and append these descriptions to the generation text prompt. This protocol gives non-in-context baselines access to the same reference information in text form, while keeping the generation inputs compatible with their original interfaces.

5.2 Evaluation Metrics

We evaluate generated videos from three complementary perspectives. First, we report **VBench** (Huang et al., 2024b; Zheng et al., 2025) metrics for general video quality, including motion smoothness, temporal flickering, and aesthetic quality. Second, we use a **VLM-as-Judge** protocol to assess physical similarity between the reference video and the generated video on a 1–5 scale. To discourage trivial reference copying, the judge is instructed to penalize generations that are visually almost identical to the reference video.

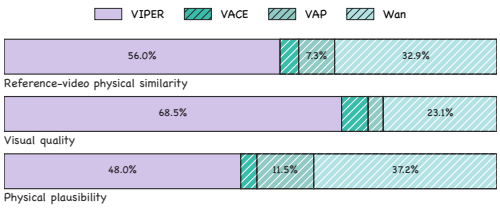


Figure 6: Visualization for human preference study.

Third, we conduct a **Human Preference** user study over reference-video physical similarity, visual quality, and physical plausibility. To keep each preference question manageable, we do not present all baseline variants simultaneously. Instead, we use Wan2.2-I2V-14B as the representative Wan-family model and VAP-Wan as the representative VAP variant, since it uses the larger backbone among the VAP baselines. Participants therefore choose among four outputs: Wan2.2-I2V-14B, VAP-Wan, VACE, and VIPER. We recruit 25 volunteers and randomly sample 10–20 comparison groups. For each group and each criterion, participants select the preferred output among the compared methods.

5.3 Quantitative Results

Table 1 shows that VIPER achieves the best aesthetic quality and the highest VLM-as-Judge physical similarity score. On motion smoothness and temporal flickering, VACE obtains the highest scores, while VIPER ranks second and remains close to the best-performing method. This trend is consistent with the qualitative comparison in Fig. 5: VACE can preserve temporal stability, but it often behaves more like pixel-level reference copying or style transfer, which weakens its ability to adapt the reference physical behavior to a new target image. In contrast, VIPER maintains competitive general video quality while achieving substantially stronger physical similarity, suggesting that reference-derived physical conditioning improves physical transfer without sacrificing the visual prior of the base I2V generator.

The human preference study further supports this trend. As shown in Fig. 6, VIPER receives higher preference rates across reference-video physical similarity, visual quality, and physical plausibility, indicating better alignment with human judgment of physically plausible video behavior.

5.4 Qualitative Results

Fig. 5 shows two reference-guided I2V results. VIPER better captures the physical behavior demonstrated by the reference video and transfers it to the target image condition. For example, in the melting case, VIPER not only generates the shape change of the ice pop but also preserves the gradual transition from transparent melted liquid to the object-specific colored fluid, which is consistent with the target material and appearance. In the deformation case, VIPER follows the physical pattern of reference video and produces a can that folds and twists from the middle region, yielding a more realistic deformation process than methods that generate only generic motion or weak shape change.

Reference videos provide crucial physical cues that short text prompts lack. While baselines maintain visual coherence, they often miss fine-grained physical dynamics. By leveraging the MLLM-to-DiT pathway to extract these reference cues, VIPER generates highly realistic physical behaviors.

5.5 Ablation Studies

To verify the specific contributions of query-based physics extraction and the hierarchical training strategy design, we compare the full model with variants that remove the learnable queries, skip individual training stages, or drop Stages 1 and 2 entirely. As shown in Table 2, the full VIPER model achieves the best overall performance among all configurations. This explicitly demonstrates that both the learnable physics queries and the hierarchical training strategy are essential for effective reference-guided physical transfer.

Table 2: Ablation study on the learnable queries and training strategy.

Variant	Motion S. ↑	Temporal F. ↑	Aesthetic Q. ↑	Physical S. ↑
w/o learnable queries	98.02	97.42	34.17	2.00
w/o Stage 1	98.51	97.08	51.01	3.41
w/o Stage 3 (LoRA)	98.49	97.24	50.57	3.33
w/o Stage 1 & Stage 2	98.45	97.03	51.05	3.26
Full	**98.60**	**97.45**	**51.86**	**3.42**

6 Discussion

VIPER shows that reference videos provide an intuitive interface for controlling physical behavior in image-to-video generation. Compared with text alone, a visual demonstration conveys dense cues about motion, contact, deformation, and material response without requiring a carefully engineered prompt. By separating target appearance from reference behavior, the learnable physics queries and hierarchical training enable the model to transfer dynamics while preserving the visual prior of the pretrained generator.

The method still depends on the physical compatibility between the reference and target. Coarse annotations may overlook motion direction, interaction strength, or temporal order, while large reference-target mismatches can lead to implausible transfer. Future work will explore finer physical annotations and pair filtering, as well as longer, multi-object, and feedback-guided generation.

AI USE STATEMENT

We used ChatGPT as an auxiliary tool during several stages of this work. Specifically, it was used to provide suggestions and feedback on dataset construction and evaluation protocols, to assist with parts of method implementation and debugging, and to support general-purpose coding, LaTeX formatting, and visualization. ChatGPT was also used for translating author-written content into English, drafting and editing parts of the manuscript, improving language and readability, and assisting with literature search and summarization.

ChatGPT was not used to formulate the core research methodology, propose or modify the main research hypotheses, interpret experimental results, derive scientific conclusions, or develop theoretical or mathematical claims. The design of VIPER, the VIPER-19K construction and filtering pipeline, the final evaluation protocol, experimental settings, and research conclusions were determined by the authors.

All AI-assisted content was reviewed and verified by the authors. AI-assisted code was tested for correctness, experiment-related suggestions were independently evaluated before adoption, and factual and bibliographic information was checked. The authors take full responsibility for the final content, implementation, experimental results, and conclusions presented in this work.

REFERENCES

Genesis Authors. Genesis: A generative and universal physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-as-prompt: Unified semantic control for video generation. *arXiv preprint arXiv:2510.20888*, 2025.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

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. In *ICML*, 2024.

Lin Geng Foo, Mark He Huang, Alexandros Lattas, Stylianos Moschoglou, Thabo Beeler, and Christian Theobalt. Physical simulator in-the-loop video generation. *arXiv preprint arXiv:2603.06408*, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025.

Nate Gillman, Yinghua Zhou, Zitian Tang, Evan Luo, Arjan Chakravarthy, Daksh Aggarwal, Michael Freeman, Charles Herrmann, and Chen Sun. Goal force: Teaching video models to accomplish physics-conditioned goals. *arXiv preprint arXiv:2601.05848*, 2026.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Chenfanfu Jiang, Craig Schroeder, Joseph Teran, Alexey Stomakhin, and Andrew Selle. The material point method for simulating continuum materials. In *Acm siggraph 2016 courses*, pp. 1–52. 2016.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Zizhang Li, Hong-Xing Yu, Wei Liu, Yin Yang, Charles Herrmann, Gordon Wetzstein, and Jiajun Wu. Wonderplay: Dynamic 3d scene generation from a single image and actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9080–9090, 2025.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Miles Macklin, Matthias Müller, and Nuttapong Chentanez. Xpbd: position-based simulation of compliant constrained dynamics. In *Proceedings of the 9th International Conference on Motion in Games*, pp. 49–54, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Ying Shen, Jerry Xiong, Tianjiao Yu, and Ismini Lourentzou. Phantom: Physics-infused video generation via joint modeling of visual and latent physical dynamics. *arXiv preprint arXiv:2604.08503*, 2026.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chen Wang, Chuhao Chen, Yiming Huang, Zhiyang Dou, Yuan Liu, Jiatao Gu, and Lingjie Liu. Physctrl: Generative physics for controllable and physics-grounded video generation. *arXiv preprint arXiv:2509.20358*, 2025a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025b.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. Univideo: Unified understanding, generation, and editing for videos. *arXiv preprint arXiv:2510.08377*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2024.

Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025.

Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025.

Haoze Zhang, Tianyu Huang, Zichen Wan, Xiaowei Jin, Hongzhi Zhang, Hui Li, and Wangmeng Zuo. Physchoreo: Physics-controllable video generation with part-aware semantic grounding. *arXiv preprint arXiv:2511.20562*, 2025a.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

A HYPERPARAMETER SETTINGS

Table 3 summarizes the hyperparameters used to train and sample VIPER. We keep the training configuration fixed across all stages unless otherwise specified in the main paper. The same sampling configuration is used for quantitative evaluation, qualitative comparison, and user study generation.

Table 3: Hyperparameter settings used for VIPER training and inference.

Hyperparameter	Setting
Batch Size	1
Accumulate Step	1
Training Strategy	ZeRO-2
Num GPU	6
Optimizer	AdamW
Learning Rate	0.00001
Learning Rate Schedule	Constant
Weight Decay	0.03
DiT Num Layers	40
Frames	81
Resolution	480x832
Connector Num Layers	3
LoRA Rank	64
LoRA Alpha	32
LoRA Layer	q, k, v, ffn.0, ffn.2
Sampler	Flow Euler
Sample Steps	50
Guide Scale	6.0

```
You are a highly rigorous Physical Attribute Classifier. Analyze the video and extract the pure physical dynamics.

[CRITICAL RULES FOR PHYSICAL REASONING - READ CAREFULLY]
1. MATERIAL DISAMBIGUATION:
   - "Rigid": Objects like glass, ceramics, wood, or metal that maintain their shape until their yield point is reached (breaking). Even if they rebound slightly upon impact, they MUST be classified as "Rigid", NOT "Elastic".
   - "Elastic": Strictly reserved for objects that significantly deform under force and return to their original shape (e.g., rubber bands, bouncy balls, springs).
2. IMPACT CLASS PRIORITY (Start vs. End State):
   - Compare the object's initial state to its final state.
   - If irreversible structural destruction occurs (e.g., breaking into pieces), you MUST prioritize "Shatter" or "Crumble".
   - "Unchanged" is STRICTLY FORBIDDEN if the object's physical structure, shape, or state is altered during the video. Structural failure > Movement.
3. FLUID vs. GRANULAR DISAMBIGUATION:
   - "Fluid": Strictly continuous liquids or gases (water, honey, smoke).
   - "Granular": Made of discrete solid macro/micro particles (sand, sugar, pebbles, pills). Even if they "flow" like a liquid, if you can perceive individual grains or if they form a conical pile at rest, it is "Granular".
4. PLASTICINE vs. ELASTIC:
   - "Plasticine": Deforms under force and STAYS deformed (e.g., clay, dough, dented metal).
   - "Elastic": Deforms under force but RETURNS to its original shape.
5. SUBJECT FOCUS RULE:
   If multiple entities interact (e.g., a hand squeezing a sponge, a hammer hitting glass), your analysis and classification MUST focus exclusively on the PASSIVE RECIPIENT of the force (the sponge, the glass). Ignore the material/category of the actor (the hand, the hammer).

You MUST output a strict JSON. For all categorical classes and confidence levels, you are FORBIDDEN from inventing new words. You MUST choose EXACTLY ONE option from the provided lists.

{
  "reasoning_scratchpad": "First, briefly describe: 1) The initial state of the object, 2) The force/action applied to it, 3) The final state. Use this state-delta to logically deduce the material and impact classes.",
  "subject_category": "Choose ONE: [Biological, Environment, Object]. (Biological = humans, animals, plants; Environment = landscapes, rooms, natural scenes; Object = inanimate items, props, materials).",
  "physics_attributes": {
    "material_class": "Choose ONE: [Granular, Fluid, Elastic, Plasticine, Inflatable, Rigid]. (Note: Glass/Ceramics = Rigid)",
    "trajectory_class": "Choose ONE: [Free-fall, Parabolic, Floating, Sliding, Rolling, Stationary]",
    "impact_class": "Choose ONE: [Shatter, Crumble, Melt, Splash, Bounce, Squish, Expansion, Shrink, Explode, Unchanged]"
  },
  "visual_features": "Texture, color, and shape of the object.",
  "synthesized_prompt": "A highly descriptive, comma-separated diffusion prompt fusing the visual features and the physical action. CRITICAL: DO NOT use the word from 'original_subject' here.",
  "flaw_analysis": "List ALL flaws (e.g., 'heavy occlusion', 'motion blur', 'out of frame'). CRITICAL RULE: If you list ANY severe flaw here that obscures the actual physical contact/change, your 'confidence_level' MUST NOT be Level_5. If no flaws, write 'None'."
  "confidence_level": "Choose EXACTLY ONE: [Level_5, Level_4, Level_3, Level_2, Level_1]. RUBRIC: \n- Level_5: The main subject clearly demonstrates distinct physical properties and dynamic changes (e.g., the collision of rigid objects, an elastic object being squeezed and bouncing, or a solid object melting).\n- Level_4: The main subject demonstrates clear physical changes, but the video has minor flaws. For example, the physical process happens too quickly, the object occupies a small portion of the screen, or there is brief occlusion, yet these do not hinder the overall observation of its physical properties.\n- Level_3: There is noticeable movement or contact, but physical deformation or state change is weak. For example, the object merely falls, slides, or translates without obvious shape changes, OR severe camera shaking and blurriness make it difficult to confidently confirm the specific physical dynamics.\n- Level_2: Mostly displays static environments or backgrounds. Even if materials like fluids (e.g., a calm lake) are present, there are no obvious, observable physical dynamic changes (such as shattering, squishing, melting, etc.).\n- Level_1: The video quality is extremely poor, or the events depicted completely violate realistic physical laws."
}
```

Figure 7: System prompt for physical video collection and attribute annotation.

B DATA ANNOTATION AND PAIR FILTERING PROMPTS

VIPER-19K uses a multimodal large language model (MLLM) for two data construction steps: annotating physical attributes and filtering reference-target pairs. Figure 7 shows the collection and annotation prompt. Instead of asking the model to describe the video in free-form language, the prompt guides it to identify observable physical cues, including material response, motion trajectory,

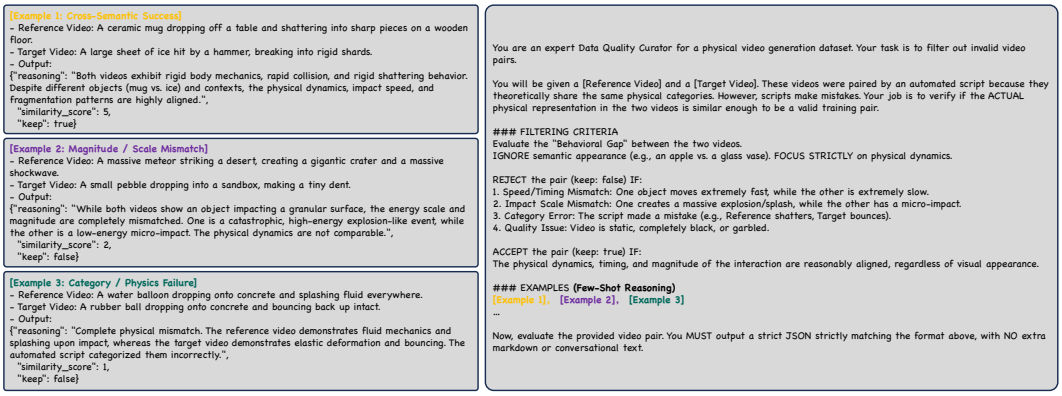


Figure 8: System prompt for reference-target pair filtering with few-shot examples.

physical impact, and visual usability. This structured annotation makes videos from different sources comparable under the same physical taxonomy and reduces the chance that later pairing is dominated by object category or scene semantics.

After annotation, rule-based pairing provides an efficient first pass by matching videos with compatible physical labels. However, these labels are still coarse relative to the requirements of reference-guided generation. Videos with the same material, trajectory, or impact label may differ in force direction, interaction strength, temporal order, contact geometry, or whether the target scene contains the physical preconditions needed for the reference behavior. As a result, purely rule-based pairs can include false positives that share high-level labels but do not support meaningful physical transfer.

Figure 8 shows the pair filtering prompt used after rule-based pairing. We add few-shot examples to make the decision boundary explicit: a valid pair should preserve transferable physical behavior rather than merely sharing object category, scene appearance, or a broad motion label. The examples also show typical negative cases, such as pairs with similar objects but incompatible contacts, pairs with matching trajectory labels but opposite motion directions, and pairs where the target lacks the required physical setup. This helps the MLLM compare the reference and target videos at the level of causal physical cues, such as material response, contact evolution, and trajectory compatibility. In practice, the few-shot prompt encourages the model to reject pairs with superficial semantic overlap and retain pairs whose dynamics can provide useful supervision for reference-guided generation.

This two-stage design balances scalability and annotation quality. Rule-based matching keeps the construction process efficient over thousands of videos, while MLLM-based filtering removes ambiguous or physically incompatible pairs before training. The resulting pairs are therefore better aligned with the goal of VIPER: using a reference video as an in-context physical demonstration rather than as an appearance template.

C Ablation on the Number of Learnable Queries

We further evaluate how the number of learnable physics query tokens affects reference-video physical conditioning. The query tokens act as a compact interface between the multimodal large language model and the video diffusion model: too few tokens may compress motion, contact, and material-response cues too aggressively, while too many tokens can introduce redundant conditioning information and extra computation. This ablation therefore examines whether VIPER relies on a sensitive query-count choice or remains stable across different query capacities.

We randomly sample 50 reference-target pairs from the test set and evaluate each variant with VBench metrics, including motion smoothness, temporal flickering, and aesthetic quality. As shown in Table 4, changing the number of queries leads to only modest differences across these metrics. This suggests that once the query bottleneck has sufficient capacity, the model can consistently extract useful physical context from the reference video. In our observation, the query count does not substantially change whether the model captures the reference physical attributes; its effect is more

Table 4: Ablation study on the number of learnable physics query tokens. We evaluate VBench metrics on 50 randomly sampled test pairs.

Num Queries	Motion S. ↑	Temporal F. ↑	Aesthetic Q. ↑
64	98.92	97.93	50.67
128	98.96	98.03	50.64
256	99.01	98.10	50.96
512	98.99	98.06	51.23

visible in general video quality. We therefore focus this ablation on VBench and do not additionally report VLM-as-Judge scores.

D Evaluation Details

We provide additional details for the VLM-as-Judge evaluation and human preference study. The VLM-as-Judge protocol evaluates whether the generated video preserves the physical behavior demonstrated by the reference video while adapting it to the target image. As shown in Figure 9, the judge is instructed to focus on transferable physical behavior, including material response, motion trajectory, contact evolution, and physical impact. We also ask the judge to penalize trivial copying of the reference appearance, because the desired output should follow the reference dynamics while matching the target condition.

```
You are a and pragmatic Physics Engine Evaluator. Your task is to evaluate the physical alignment between two videos: a Source Video (Reference) and a Generated AI Video (Target).

CRITICAL EVALUATION MINDSET:
Current AI video generation models inherently produce minor visual artifacts, flickering, slight morphing, or imperfect physics. You must be HIGHLY LENIENT towards these expected generative flaws. The default expectation should be a positive score (3, 4, or 5). Reserve scores of 1 or 2 ONLY for complete physical nonsense or literal pixel-level clones.

### EVALUATION PROCESS (Chain-of-Thought)
STAGE 1: REFERENCE ABSTRACTION
Observe the [Video 1] Reference Video. Extract its pure physical dynamics into text. Ignore colors, lighting, background, and the specific identity of the object. Focus ONLY on:
- Material properties (rigid, fluid, elastic, etc.).
- Impact/Deformation behavior (shattering, bouncing, etc.).
- Motion/Trajectory.

STAGE 2: CLONE DETECTION & TARGET COMPARISON
Observe the [Video 2] Target Video.
1. Clone Detection: Is it a literal pixel-for-pixel copy? (Note: Sharing a similar color palette or layout is perfectly fine. Only penalize if it is identical).
2. Physics Comparison: Focus on the overarching physical INTENT rather than micro-flaws. If the Target is supposed to show a bouncy ball, does it bounce? Overlook minor clipping, unnatural stretching, or brief disappearing pieces, as long as the general physical rule is attempted.

### SCORING RUBRIC (Lenient Standard)
- 5 (Excellent): Target successfully replicates the core physical dynamics extracted in Stage 1 and applies it to a new scene/object. Generative artifacts (flickering, minor morphing) are completely acceptable and should not prevent a score of 5.
- 4 (Good): The overarching physical rules match the reference. There may be noticeable physical approximations (e.g., an object breaks slightly incorrectly, or stretches when it shouldn't), but the main physical behavior is clearly aligned. This should be a very common score.
- 3 (Passable): The model clearly attempted to match the physics, but struggled. There are obvious physical flaws (e.g., a fluid moves a bit like a solid, or severe clipping occurs upon impact), BUT the basic concept from the reference is still recognizable.
- 2 (Poor): Major fundamental physical failure. The behavior contradicts the reference (e.g., a glass should shatter, but it melts instead).
- 1 (Catastrophic/Clone): Total collapse into pure noise, absolute disregard for the physics, OR a lazy, exact pixel-level clone of the Reference.

### OUTPUT FORMAT
You MUST output your evaluation in the following strict JSON format:
{
  "step1_reference_physics": "Extracted physical rules from the Reference Video...",
  "step2_clone_detection_and_comparison": "Assessment of whether it is a true visual clone, followed by a forgiving comparison of physics...",
  "is_visual_clone": <boolean true/false>,
  "physical_alignment_score": <int 1-5>
}
```

Figure 9: System prompt of VLM-as-Judge.

The human preference study asks participants to compare generated videos along reference-video physical similarity, visual quality, and physical plausibility. We build the preference interface with Gradio, as shown in Figure 10. Each question presents the same reference-target condition and outputs from the compared methods, so participants can judge whether a method transfers the reference behavior without ignoring target appearance. We recruit 25 volunteers and randomly sample 15–20 questions for each participant. This design keeps each comparison manageable while collecting judgments that directly correspond to the main evaluation criteria.

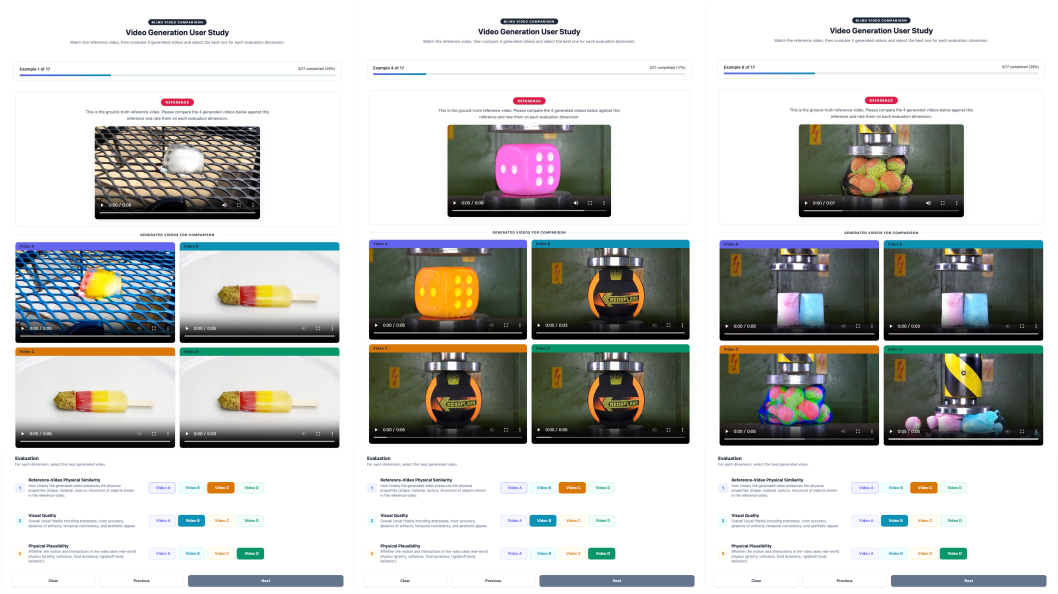


Figure 10: Gradio interface used for the human preference study. Participants compare generated videos under the same reference-target condition and vote according to the specified criterion.

E Additional Results

Figure 11 shows that VIPER can also use reference videos that are not captured from the real world. Even when the reference video is synthetic or visually unrealistic, it can still contain transferable physical cues, such as material response, motion trajectory, and interaction pattern. VIPER extracts these physical attributes from the reference video and transfers them to the target image, suggesting that the reference video serves as a visual in-context demonstration rather than merely a source of real-world appearance.

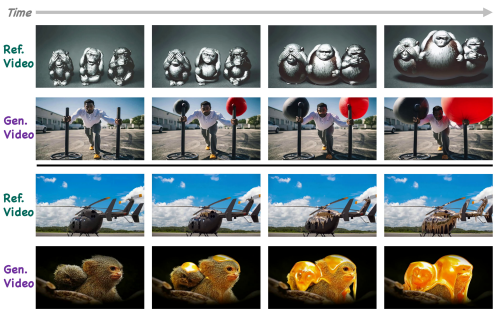


Figure 11: Reference-guided generation with non-real-world reference videos. VIPER can extract transferable physical behavior from synthetic or visually unrealistic references and apply it to the target condition.

Figure 13 provides more qualitative examples across diverse reference-target pairs. These results further illustrate that VIPER can transfer reference-derived physical behavior while adapting the generated content to different target images. We place the full qualitative sheet at the end of the supplementary material to keep the explanatory text and compact figures together in the preceding pages.

F LIMITATIONS AND FAILURE CASES

Figure 12 shows representative failure cases of VIPER. These cases suggest that reference-guided physical transfer can still be limited by coarse data annotations and challenging reference-target mismatches.

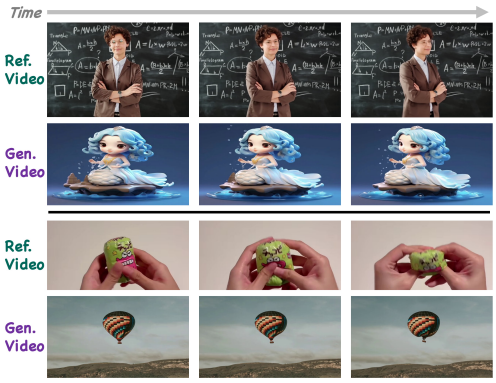


Figure 12: Failure cases of VIPER, including trajectory-direction ambiguity and over-transfer under mismatched physical context.

The first failure mode is caused by insufficiently fine-grained physical labels. For example, a single "Spin" label does not distinguish clockwise and counterclockwise rotation, so rule-based pair construction may group videos with different motion directions. The second failure mode occurs when the reference and target conditions differ too strongly in physical context. In such cases, VIPER may correctly identify the reference behavior but transfer it to a target scene where the required force or interaction is absent. Future work can reduce these errors with finer physical taxonomies, direction-aware trajectory annotations, and stronger filtering for reference-target physical compatibility.



Figure 13: Additional qualitative results of VIPER on diverse reference-target physical transfer cases.

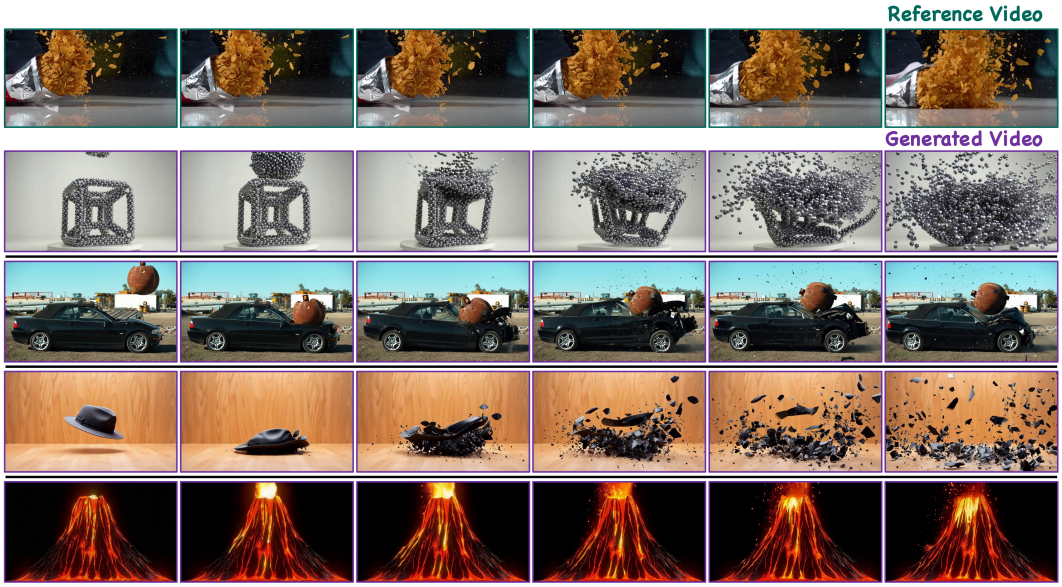


Figure 14: Additional results of different target images guided by the same reference video.



Figure 15: Additional results of different target images guided by the same reference video.



Figure 16: Additional results on diverse reference-target physical transfer cases.