Physics Question Scene Graph: Fine-grained Evaluation of Physical Plausibility in Text-to-Video Generation

Atin Pothiraj[1], Jaemin Cho[2,3], Yue Zhang[1], Elias Stengel-Eskin[4], and Mohit Bansal[1]

[1] University of North Carolina at Chapel Hill, USA
[2] AI2, USA
[3] Johns Hopkins University, USA
[4] University of Texas at Austin, USA
https://github.com/atinpothiraj/pqsg

Abstract. Video generation models are increasingly capable of producing realistic videos, but they still struggle to generate videos that follow basic physical laws. Compounding this is a lack of reliable granular evaluation methods for localizing and specifying physical law violations in videos. We address this by introducing Physics Question Scene Graph (PQSG), a hierarchical question-based evaluation pipeline. PQSG evaluates generated videos by checking their faithfulness to a prompt across objects, actions, and adherence to physical laws using a graph-based hierarchy of questions generated by a vision-language model (VLM), guided by high-quality in-context examples. By representing questions as a graph, PQSG introduces logical dependencies within questions, ensuring that each query is contextually valid. Moreover, PQSG provides granular assessments of which qualities of the video violate physical plausibility constraints. We validate PQSG by creating FinePhyEval, a dataset with physics-based prompts and corresponding generated videos from diverse state-of-the-art video generation models (Sora 2, Veo 3, and Wan 2.1), with each video annotated across multiple categories by humans. Using FinePhyEval, we measure the correlation between PQSG's fine-grained scores and human judgments, showing higher overall correlations than prior work. We also find that PQSG ranks closed-source models higher than Wan 2.1 on physical realism. Lastly, we show that the annotations we provide in FinePhyEval can also be used for subtask evaluation: we benchmark two strong VLMs on generating and answering questions, finding that while models can create human-like questions, they still fall short of human performance in answering them.

Keywords: Physics · Evaluation · Video · Text-to-Video Generation

1 Introduction

An intuitive understanding of physical laws has been posited as a core component of human cognition [24], allowing us to reason accurately about the world

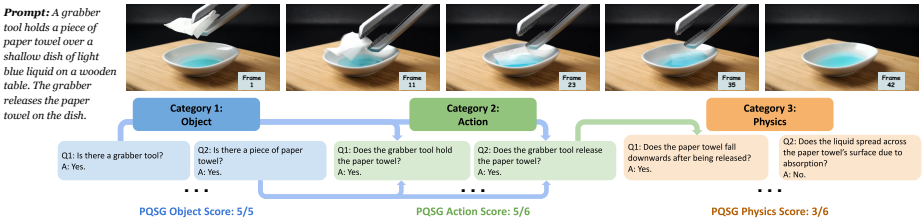

Fig. 1: An example prompt and its corresponding generated video from Wan 2.1, with sample PQSG nodes and edges (not all are included) for each category. While the video contains the correct objects (a grabber, paper towel, bowl, etc.) it does not show all the right actions, and the physical interactions shown are implausible, with the paper towel dissolving into the liquid rather than absorbing it. PQSG specifies in which category the video is unrealistic (in this case, in its physics) and, within that category, which physical interactions are implausible (here, that the liquid fails to absorb into the paper towel, reflected in Physics Q2).

and the consequences of our actions in it. This capacity, often called "world modeling," has long been a goal of video generation models [35, 38]: to reason about future world states by rendering future video frames. Indeed, recent video generation models have become increasingly adept at generating photorealistic videos of objects and people in various environments [5]. However, while some understanding of basic physical laws has been documented even in infants [42, 43], current video generation models still struggle with physical realism, producing outputs that violate fundamental physical laws, such as solid mechanics, fluid dynamics, and optics [33]. Adherence to these laws is key for developing true world models that can serve as practical simulators. For example, a simulation of dish-washing must accurately represent water flow and the contact between a sponge and a dish. Such physical accuracy is critical for downstream applications in robotics [9], embodied agent training [41], and synthetic data generation [7].

Compounding current models' shortcomings in producing physically-plausible videos is a lack of reliable, fine-grained, and automated methods for evaluating physics adherence. Existing methods provide only high-level, coarse-grained scores, failing to capture the inherent structure of physics evaluation. We argue that a robust physics evaluation must capture the logical dependencies within a scene. For example, to evaluate whether an object's *fall* is physically plausible, one must verify that the object is *present* (object-level correctness) and that it is *falling* (action-level correctness). Only then can the *physics* of the fall (e.g., acceleration) be assessed. Existing methods (e.g., [17, 23, 33]) lump these factors into a single, aggregate score, making it difficult to determine *why* a video failed. Moreover, these combined metrics are often fooled by videos that are visually realistic and temporally consistent, but may still be physically implausible [54]. For example, in Fig. 1, baseline methods may assign a high score to a video where human annotators identify a clear physics violation (e.g., a paper towel dissolving instead of absorbing liquid). As models become more powerful, the errors they display will also become subtler, motivating the need for a fine-grained

evaluation metric that can provide feedback on specific physical inaccuracies, both for evaluation and for fine-grained video repair [25].

To address this gap, we introduce the **P**hysics **Q**uestion **S**cene **G**raph (PQSG), a novel, structured evaluation framework for video generation models that measures fine-grained details of physical dynamics. PQSG is designed with two key properties: (1) PQSG is **granular**: rather than assessing adherence to physical laws in one, overarching score, it decomposes the assessment into multiple sub-queries, allowing for granular feedback on exactly which aspect of the video violates a physical law. (2) PQSG follows a **dependency structure**, ensuring that questions are asked in a sensible way: if an object is not present in a video, we do not subsequently ask about actions or physics interactions involving it. We illustrate this in Fig. 1, where the sub-queries are divided into three hierarchical categories: **object**, **action**, and **physics**. Note that PQSG evaluates the semantics *mentioned in text prompts* — it assesses whether a generated video realizes the objects, actions, and physical interactions specified by its prompt. We formulate PQSG as a two-stage process: first, in the question-generation stage (QG), we task vision-language models (VLMs) with generating hierarchical physics-aware graphs of questions from text prompts. In the second stage – question-answering (QA) – these physics scene graphs are then passed with the generated video to a VLM to answer questions about the video, producing interpretable object, action, and physics scores, as seen in Fig. 1 (as well as Fig. 2). In the case of Fig. 1, PQSG's QG stage generates targeted questions (e.g., questions about fluid dynamics) which are then answered, leading to appropriately low physics scores.

To evaluate the components of PQSG (QG, QA) and subsequently evaluate diverse video generation models using PQSG, we collect FinePhyEval, a dataset of 195 human-annotated prompt-video pairs, with prompts sourced from Physics-IQ [33] and videos generated from three diverse state-of-the-art video generation models: Sora 2 [37], Veo 3 [12], and Wan 2.1 [45]. FinePhyEval provides comprehensive human annotations for: (1) video quality across four Likert-scale categories (object, action, physics, and overall quality), where we find strong inter-annotator agreement; and (2) PQSG's constituent subtasks of question-generation (QG) and question-answering (QA).

On FinePhyEval, we first show that PQSG results in better correlations to overall human ratings than prior metrics, and that – unlike prior metrics – PQSG can provide fine-grained evaluation feedback at the level of individual attribute categories. Moreover, we rank different models according to PQSG's automated metric, finding that proprietary models (Sora 2, Veo 3) outperform the open-source models (Wan 2.1, Cosmos-14B), and that all models struggle on action and physics more than object generation. We also ablate the components of PQSG, measuring the contribution of fine-grained questions and dependency graphs. Additionally, the annotations we provide in FinePhyEval allow for low-level subtask evaluation on QA and QG: we benchmark two strong VLMs – Gemini-2.5-Pro [10] and GPT-5.5 [36] – on their ability to perform QG and QA, finding models to be adequate at QG but well short of human performance

on QA despite the relatively short length of videos (avg. length: 4.39 seconds). In the overall evaluation, we find that for QA, action and physics categories pose challenges to models, with the best model, GPT-5.5, obtaining $\sim 65\%$ QA accuracy. This suggests promising directions for improvement in evaluation: we show that PQSG has a high upper bound correlation with human judgments when QA is performed by a human, suggesting further gains will be obtained as VLMs improve.

2 Related Work

Video Generation Models. Progress in text-to-video generation has accelerated rapidly with the advent of video diffusion models [4,18–20,34,50,51,53]. Diffusion models can now realistically render complex high-resolution scenes, with recent examples approaching the realism of human-shot videos [12,37]. However, recent studies indicate that even state-of-the-art models frequently violate basic physical principles [3, 33]. To address this, previous research has proposed various physics-aware enhancements, including motion guided by the physics-simulator [28,32,44], physics-informed post-training and reward optimization [21, 26, 27, 49], implicit physical priors injected through vision-language reasoning or force-based conditioning [11, 14, 22, 48, 52], and program-based generative models that explicitly encode physical laws [28]. Critically, making progress towards more physically-realistic models requires a fine-grained evaluation framework that offers precise, actionable metrics on *which* physical laws and phenomena current models struggle with.

Evaluation of Generated Videos. Early evaluation metrics for image and video generation have focused primarily on general visual quality or semantic alignment with textual prompts, employing methods based on deep feature embeddings [17, 47], classifier predictions [40], semantic matching [16], or comprehensive evaluation toolkits [29]. However, these metrics do not explicitly measure physical realism or pinpoint precise violations of physical laws, as their criteria inherently overlook fine-grained physical constraints and nuanced action semantics. Recent evaluation frameworks for video generation have attempted to address physical realism, prompt alignment, and factual consistency through several approaches. Methods such as VideoScore [15] primarily evaluate semantic and factual alignment, but inherently overlook subtle yet critical violations of physical laws. More specialized methods, such as VideoPhy-2 [3], try to mitigate this gap by fine-tuning vision-language models (VLMs) to produce semantically meaningful realism scores. In a similar vein, PhyGenEval [31] leverages VLMs for assessing physical commonsense. Our work differs from prior efforts in its focus on structured, fine-grained evaluation. Unlike prior work that provides a single aggregate score or a few coarse-grained scores, PQSG generates a hierarchical set of detailed questions that localize exactly where a generated video should be improved. This granular QA format makes PQSG easily interpretable and ensures that evaluations are consistent across answerers, human-aligned, and generalizable. In a similar spirit to our work, DSG [6] offers question-based eval-

uation using graphs. However, DSG is designed for images and fails to test for temporal properties (such as actions or physics-based interaction), as shown in our evaluation (see Sec. 5.1).

3 Physics Question Scene Graph (PQSG)

Given a text prompt and a generated video, we aim to output video scores across multiple dimensions with a special focus on adherence to physical laws and with the goal of developing an evaluation system that identifies the specific failure points in generated videos, enabling fine-grained analysis. For example, consider the prompt: *"Two pillows on a table and two grabber tools hanging above them from which a brown tennis ball and an orange block are suspended. The grabber tools let go of the ball and block."* One key physics-related moment is the tennis ball falling, and we want to examine whether the fall seems plausible in accordance with the laws of gravity. However, evaluating this "plausible fall" presents two significant challenges. On one hand, many general video evaluation frameworks focus on foundational elements like object existence (e.g., "Is there a ball?") and attribute binding (e.g., "Is the ball brown?"). These methods are often insufficient for evaluating complex object dynamics – the core of a physics evaluation. On the other hand, for a framework to successfully judge the dynamics of the fall, it must first verify its preconditions: (a) that the tennis ball exists, and (b) that the ball is being dropped. If either of these preconditions fails, the video generation model has failed before even attempting the physics, rendering a judgment on the "fall's realism" impossible and making such judgments potentially hallucinated or misleading. Previous frameworks have generally failed to model this specialization in dynamics and these prerequisite dependencies. To jointly address both challenges, we introduce PQSG.

The **Physics Question Scene Graph (PQSG)** is a framework for fine-grained evaluation of a generated video by defining the physical scene as a directed acyclic graph (DAG). Following DSG [6], this graph consists of atomic verification questions (nodes) with explicit logical dependency structures (edges), all generated from the text prompt describing the scene. We show an example video evaluation with PQSG in Fig. 2.

Nodes. PQSG organizes nodes into three hierarchical categories that can produce separate, fine-grained scores:

- **Object**: Verifies that a key object from the prompt is present in the video (e.g., *"O2: Are there two pillows?"*).
- **Action**: Verifies the object is exhibiting the correct action (e.g., *"A1: Does one grabber tool release the brown tennis ball?"*).
- **Physics**: Evaluates the plausibility of actions regarding physical laws (e.g., *"P8: Do the pillows visibly deform or compress upon impact?"*). In general, actions and physics are distinguished by whether they are mentioned in the prompt: actions are more directly tied to the prompt, whereas physics interactions (e.g., deformation, absorption, etc.) are implicit and part of physical commonsense.

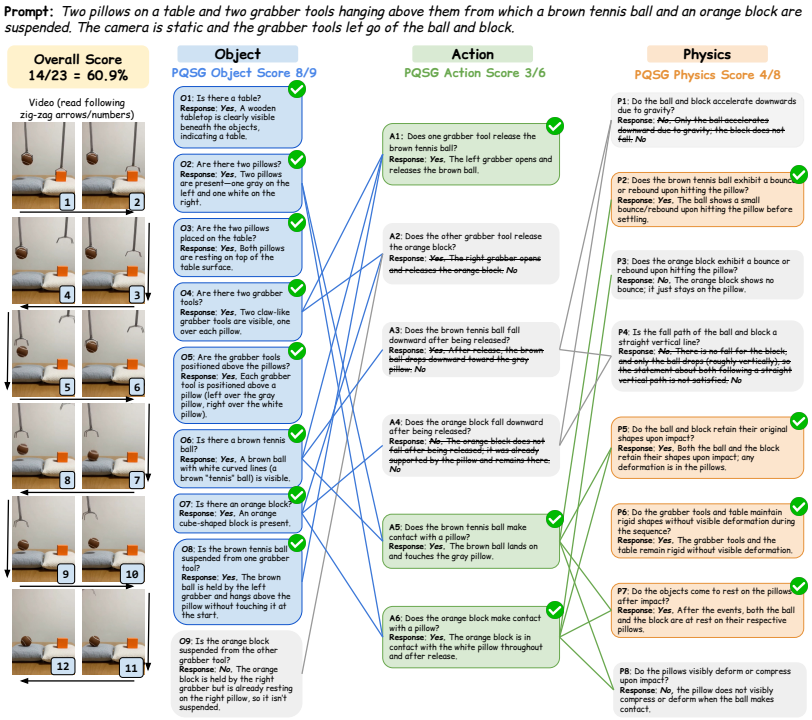

Fig. 2: Illustration of fine-grained evaluation on a generated video with PQSG. For each question (e.g., P8: *"Do the pillows visibly deform or compress upon impact?"*), PQSG provides binary judgment and reasoning behind each judgment (*"A: No, the pillow does not visibly compress or deform when the ball makes contact"*). When a parent question is answered *"no,"* then its children's questions are automatically also marked as *"no"* (e.g., node A2 becomes invalid because its parent node, O9, is answered with *"no"*). From the per-question evaluation result, we obtain per-category scores and an overall score, which is the sum of the per-question scores. The fine-grained questions pinpoint the detailed failure modes of generated videos.

This explicit separation of **Action** node from **Physics** node is what allows our method to pinpoint failures by isolating physical plausibility from simpler generation failures: (a) a video fails to show an action in the first place *vs.* (e.g., an object does not fall) (b) a video shows an action, but the action may not look plausible regarding physical laws (e.g., an unrealistic fall).

Edges. PQSG organizes questions into a hierarchical graph encoding their dependencies. As illustrated in Fig. 2, most of the action binding nodes have object existence nodes as parents, and most of the physics nodes have action binding nodes as parents. This hierarchy ensures that an object must exist and undergo the correct action before its physics can be judged. Beyond these cross-level dependencies, PQSG also admits same-category edges like Object-to-Object, Action-to-Action, and Physics-to-Physics, which capture sequential dependen-

cies within a category, such as a later action that presupposes an earlier one, or a physical outcome that is contingent on a preceding physical state. Edges are thus drawn from object to action, from action to physics, or between nodes of the same category, and no other connections are permitted. During evaluation, this dependency structure is strictly enforced. If the answer to a parent node's question is "no" (e.g., the Object node fails), the questions from its entire chain of child nodes are not queried and are automatically also marked with "no." A child question is only queried if all its prerequisite parent nodes have been answered affirmatively. This ensures that only answerable questions are actively posed to the model, reducing the potential for hallucination. As shown in our ablation study (Table 8), this dependency structure yields stronger human correlation than unstructured approaches.

Evaluation Pipeline. Our evaluation pipeline consists of two steps: Question Generation (QG) and Question Answering (QA). Both naturally lend themselves to VLM implementations (e.g., via Gemini 2.5 Pro [10], GPT-5.5 [36]).

In the QG step, we prompt the VLM with task instruction and one in-context example of a prompt, PQSG nodes, and PQSG edges, followed by a new prompt we want to generate PQSG from. See appendix for the QG prompt. In the QA step, we prompt a VLM with a generated video and a PQSG node (a question), and let the model answer one question at a time. In our initial experiments, we find that letting the QA model answer only with yes/no responses often discourage the use of chain-of-thought reasoning. Instead, we implement QA in two steps: (a) generating an open-ended response and (b) categorizing the answer as "yes" or "no," using the same VLM for both steps. We find this strategy improves the QA quality while adding negligible latency. After marking responses from invalid nodes as "no," we take the ratio of the "yes" response as the average score. In addition to the average score, we can also calculate separate, fine-grained scores for each of the three node categories (Object, Action, and Physics).

In Fig. 2 we illustrate an evaluation of a generated video with PQSG. PQSG provides per-question level fine-grained evaluation in binary judgment and reasoning behind each judgment. From the per-question evaluation result, we can obtain per-category scores and an overall score, which is the sum of all per-question scores. Furthermore, the fine-grained questions help pinpoint the detailed failure modes of generated videos.

4 FinePhyEval: Human Annotation of Fine-grained Video Scores

To analyze the utility of fine-grained video evaluation by PQSG, we collect FinePhyEval, a new dataset of human judgments across fine-grained categories for videos generated by recent SoTA video generation models.

4.1 Prompt and Video Generation

Evaluation Prompts & Reference Videos. We source our prompts from Physics-IQ [33] because it contains prompts designed to test video generation

models' understanding of physical principles and also provides a reference ground-truth video for each prompt. We utilize all 65 prompts from the dataset; an example prompt is shown in Fig. 2, highlighting the prompts' complexity and compositional nature, with multiple objects and object interactions.

Video Generation Models. While the original Physics-IQ paper reports the performance of older video generation models (e.g., Sora, Runway, Stable Video Diffusion), recent models demonstrate markedly improved capabilities. To validate our evaluation system with the most recent and most visually realistic models, we collect videos from three recent state-of-the-art models, Sora 2 [37], Veo 3 [12], and Wan 2.1 [45], based on their high rankings on the Artificial Analysis text-to-video leaderboard [2] (as of March 2026) and their public API/checkpoint availability. We additionally evaluate Cosmos-Predict2.5-14B [1], a model purpose-built as a physical world simulator, to test whether an architecture designed for physical realism yields more physically plausible videos under our fine-grained evaluation. We generate 260 prompt-video pairs (65 per model) using their default configurations. This yields 4-second, 720x1080 (30 fps) videos from Sora 2; 4-second, 1280x720 (24 fps) videos from Veo 3; 5-second, 1280x720 (16 fps) videos from Wan 2.1; and 6-second, 1280×704 (16 fps) videos from Cosmos-Predict2.5-14B.

4.2 Human Annotation of Question Generation and Answering

Human Annotation of Fine-grained Verification Questions. To establish a ground-truth set of verification questions – i.e., to establish ground-truth question generation (QG) – we manually annotate verification questions for a sample of 20 prompts from FinePhyEval. Following DSG [6], we ensure that the questions cover the full content from the prompt, while each question is atomic (asking only one aspect of the prompt at a time) and does not overlap with another question. We later measure the precision and recall of automatically-generated questions from the QG system against these questions.

Human Annotation of Answers to Verification Questions. To establish an upper-bound for question-answering (QA) performance, we collect human yes/no answers to the generated questions on 30 prompt-video pairs sampled for the QG verification step. In total, we manually annotate 444 QA pairs, as each prompt-video pair has a PQSG with multiple questions in it.

4.3 Likert-scale Video Judgments

Human Annotation of Likert-scale Video Scores across Four Categories. To validate the overall scores produced by different methods, we collect human Likert score judgments over generated videos. Specifically, for each of 195 videos in FinePhyEval, we collect four text-video alignment scores (object, action, physics, and overall categories), with eight non-author human annotators; i.e., $4 \times 195 = 780$ scores in total. Each score is measured in Likert-scale (1-5). Object, action, and physics categories measure object existence, high-level action

Table 1: Correlation between human overall Likert scores and different metrics on FinePhyEval.

Metric	Pearson's r	Kendall's τ	Spearman's ρ
VideoScore [15]	0.289	0.262	0.378
VideoPhy-2-Autoeval [3]	0.346	0.277	0.349
PhyGenEval [31]	0.272	0.220	0.264
DSG [6]	0.302	0.195	0.265
Direct VQA	0.382	0.290	0.360
PQSG (Ours)			
w/ Gemini-2.5-Pro	0.467	0.306	0.406
w/ GPT-5.5	**0.478**	**0.336**	**0.456**

faithfulness, and physical plausibility, respectively. The overall category measures an annotator's judgment of video-text alignment without focusing on the three criteria. We study how this overall judgment Likert score is correlated with different video evaluation metrics (Table 1) and category-specific human Likert scores (Table 2). Note that this four-category human judgment for text-video alignment is more fine-grained than many previous works that asks annotators to consider one or two categories; we study correlation between per-category human Likert score and per-category PQSG score in Table 3. We also calculate the intraclass correlation coefficient (ICC) on a subset of FinePhyEval and find a high inter-annotator agreement of 0.84 across categories (see appendix). See appendix for annotation details, including full annotator guidelines.

5 Experiments and Discussion

We first compare our PQSG with other single summary score methods in terms of human judgments (Sec. 5.1). Then we show the comparison of video generation models in PQSG scores (Sec. 5.2). We also analyze the reliability of two subtasks of PQSG: question generation and question answering (Sec. 5.3). Moreover, we demonstrate PQSG generalizes to an external dataset and an open-source VLM (Sec. 5.4). Lastly, we provide ablation analysis of PQSG design choices (Sec. 5.6). See the appendix for more qualitative examples and ablation studies.

5.1 Correlation to Human Judgments: Generated Video Evaluation

Correlation with Overall Human Likert Scores. While the focus of PQSG is to provide detailed failure mode analyses instead of providing a single scoring method, we note that its scores can be aggregated into an overall score. Here, we experiment with PQSG as an overall video scorer for use cases where an aggregate score is desired. We compare PQSG with recent video evaluation metrics for physical plausibility: VideoScore [15], VideoPhy-2-Autoeval [3], and

PhyGenEval [31]. We include DSG [6], an evaluation for image generation models, as a baseline. For fair comparison with DSG, we use Gemini-2.5-Pro with DSG's QG/QA model components, and feed the whole video as input to the QA component, like ours. We also include a simple direct VQA baseline; i.e., asking Gemini 2.5 Pro to output an alignment score between 1 and 5, given a video and a text prompt. We evaluate Pearson's r, Kendall's τ, and Spearman's ρ. Table 1 shows that PQSG achieves similar or better correlation with human judgment in all three correlation coefficients.

What do human evaluators care about when measuring "overall" score? When human evaluators are asked to judge a video with a single overall score, they might not judge videos based on an unweighted average of three categories; in other words, they may consider some categories more or less strongly in forming their final judgment (or indeed, may consider alternate factors not considered in our three categories). We further explore this in Table 2, where we study the correlation between different human Likert scores for object/action/physics categories and their overall rating for the video as a whole (see Sec. 4). Physics correlates most strongly with the overall score ($r = 0.85$), indicating that the annotators' final assessments are more predicted by errors in physics than other categories. This underscores the importance of correct physics evaluation.

Table 2: Pearson's r correlation between different human Likert score categories (per-category to overall).

	Object	Action	Physics
Overall	0.44	0.66	0.85

Per-category PQSG-to-human correlation. We study category-specific correlations between PQSG's three category-specific scores and human's category-specific Likert scores. For PQSG, we report two versions: one using answers from GPT, and one using human annotators. Note that the Likert score and fine-grained question-specific evaluation have different purposes: the former is intended as an overall summary score for ranking, while the latter is designed to provide detailed feedback, with the score itself being aggregated from the answers to many granular questions. Because of this, we expect to see a moderate positive correlation (> 0.4) instead of a very high correlation ($0.9 >$). Table 3 shows this correlation in all three categories. Among categories, automated QA performs nearly identically to human scores in object categories, while there is some room to improve in action and physics categories. This is also consistent with our findings in QA evaluation (Table 5).

5.2 Comparing Video Generation Models via PQSG

In Table 4, we compare Sora 2, Veo 3, Wan 2.1, and Cosmos-14B on FinePhyEval prompts across the three granular categories of our metric. Here, we use automated QG and QA on the whole FinePhyEval. Overall, we find that models struggle on action and physics, with lower scores than for object prediction. This indicates that while models tend to generate the correct objects

Table 3: Pearson's r correlation between different per-category human Likert scores and per-category PQSG scores (with both VLM and human QA).

	Human Likert category - PQSG category		
	Object-Object	Action-Action	Physics-Physics
w/ GPT-5.5 QA	0.59	0.68	0.48
w/ Human QA	0.59	0.73	0.57

Table 4: Comparison of different video generation models. PQSG scores are averaged over three question-generation runs.

Video Gen.	Object	Action	Physics	Overall
Sora 2	$0.95_{\pm 0.042}$	$0.75_{\pm 0.056}$	$0.69_{\pm 0.070}$	$0.78_{\pm 0.057}$
Veo 3	$0.98_{\pm 0.016}$	$0.78_{\pm 0.028}$	$0.68_{\pm 0.022}$	$0.80_{\pm 0.018}$
Wan 2.1	$0.86_{\pm 0.029}$	$0.53_{\pm 0.058}$	$0.46_{\pm 0.035}$	$0.59_{\pm 0.042}$
Cosmos 2.5	$0.93_{\pm 0.040}$	$0.56_{\pm 0.057}$	$0.46_{\pm 0.078}$	$0.62_{\pm 0.038}$
Average	0.93	0.66	0.57	0.70

Table 5: Human evaluation of question answering on 30 FinePhyEval videos, measured with accuracy.

QA Method	Accuracy (%)		
	Object	Action	Physics
Gemini-2.5-Pro	87.6	59.5	61.5
GPT-5.5	**88.4**	**63.4**	**64.6**

(some with very high accuracy), they struggle with producing the right actions and physical interactions in the video. Moreover, proprietary models (Sora 2 and Veo 3) outperform the open-source variants (Wan 2.1 and Cosmos-14B) by a large margin. The narrow confidence intervals confirm that PQSG's model rankings remain consistent across repeated question-generation runs.

5.3 Subtask Evaluation

As FinePhyEval provides fine-grained human annotation of questions and answers, this allows the detailed evaluation of PQSG in two subtasks: question-generation (QG) and question-answering (QA). Here, we manually annotate 20 graphs to compare VLMs to human-level graph generation, and answer generated questions to provide an upper-bound to the metric with human-level answering.

Question Generation. We evaluate generated PQSG questions in terms of how accurate they are (precision) and how completely they cover the different aspects of the detailed prompts used (recall). Concretely, given a set of human-annotated questions $Q^h = q_1^h, ..., q_{|Q^h|}^h$, and generated questions $Q^g = q_1^g, ..., q_{|Q^g|}^g$, let $m_i \in 0, 1$ be 1 if q_i^h is semantically covered by any one or multiple generated questions from Q^g and 0 otherwise, and similarly, let $m_j \in 0, 1$ be 1 if q_j^g is semantically covered by any one or multiple human annotated questions from Q^h and 0 otherwise. We manually calculate **precision** $= \sum_{j=1}^{|Q^g|} m_j / |Q^g|$ and **recall** $= \sum_{i=1}^{|Q^h|} m_i / |Q^h|$. We compare GT questions described in Sec. 4 with questions generated with two VLMs: Gemini-2.5-Pro [10] and GPT-5.5 [36].

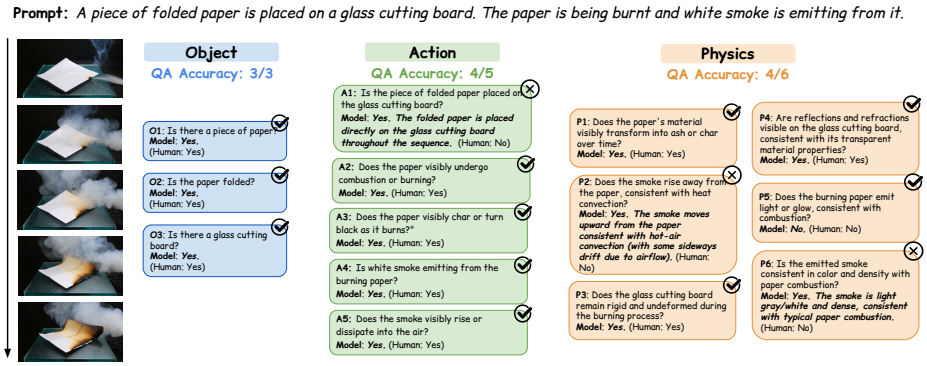

Fig. 3: An example video where a QA model (GPT-5) struggles. The model fails to capture the complex dynamics of the smoke, exhibiting a strong "yes-bias" [39, 46] by defaulting answers to "yes," while human answers include multiple "no" responses.

Table 6: Human evaluation of question generation on 20 prompts.

QG Method	Precision (%)	Recall (%)
Gemini-2.5-Pro	**95.2**	**95.2**
GPT-5.5	92.0	**99.6**

Table 7: Pearson's r correlation with human judgment on VideoPhy-2.

Metric	SA	PC
VideoPhy-2-Autoeval	0.450	0.420
+ PQSG (Ours)	**0.550**	**0.498**

VLMs can generate highly reliable verification questions. As shown in Table 6, PQSG questions generated by both VLMs (Gemini-2.5-Pro and GPT-5.5) are highly aligned with human-annotated questions, in terms of how accurate they are (precision) and how completely it covers the aspects of the detailed prompt (recall). When examining manual annotations, the mismatches we observe are mostly due to a failure to predict future states of a prompt (such as not accounting for a possible state where objects other than the main subject of the prompt could interact). Even with this minor failure, generating PQSGs with VLMs is reliable with precision and recall scores above 90%.

Question Answering. Given properly generated questions, a good question answering (QA) model is expected to answer the verification question in a way that aligns with human answers. We examine how accurately different QA models answer each question by comparing the yes/no responses generated by the models with ground-truth answers derived from human annotations (see Sec. 4). The comparison is evaluated using accuracy as the primary metric. Building on the finding in Table 6 that models are strong question generators, we use automatically-generated questions from Gemini-2.5-Pro. We then evaluate Gemini-2.5-Pro and GPT-5.5 as QA models. For each model, we process videos at their maximum frames-per-second (fps): 24 fps for Gemini-2.5-Pro and 8 fps for GPT-5.5.

VLMs perform well at evaluating objects and high-level actions, but struggle with physical plausibility. Table 5 shows the accuracy of Gemini-2.5-Pro and GPT-5.5 models in question answering for three categories. For object and action categories, the answer accuracy is high, but for physics, the answer accuracy is low. This aligns with recent findings that VLMs struggle with physical reasoning [8] and spatio-temporal reasoning [55]. Manual examination of model predictions reveals common failure modes in identifying rapid actions and in capturing complex physical interactions (e.g., the direction and density of smoke from a flame). To further analyze the failure modes of VLMs on QA, we provide a qualitative example in Fig. 3. Here, we find that the model generally suffers from "yes-bias" [39, 46], where it tends to answer yes to questions. The mistakes the model (in this case, GPT-5.5) makes are also reflected in its reasoning for incorrect predictions, where it confidently claims that certain actions have occurred even though they are not present in the video. Part of the failure on physical questions may stem from the VLM's LLM component and the commonsense knowledge embedded in it. For example, it suggests the smoke is rising upward with a slight sideways drift due to airflow, whereas the video actually shows smoke being ejected sideways, similar to a flare. with Taken together, these results suggest that verifying object and high-level actions in videos can be automated in high precision using recent VLMs. However, there remains significant room for improvement in evaluating low-level, rapid actions as well as assessing the physical plausibility of events in videos.

5.4 Generalization Study: External Dataset and Open-source VLM

We demonstrate PQSG's improved performance on 100 prompt-video pairs from the VideoPhy2 [3] test set using the same in-context examples and prompts as other experiments. Here, we show the generalization capability of PQSG in a different setup. In Table 7, PQSG increases both semantic adherence (SA) and physical commonsense (PC), showing the generalization of PQSG in another dataset and with an open-source VLM.

5.5 Iterative Refinement with PQSG

We investigate whether PQSG can be used to directly improve video generation. Because PQSG decomposes quality into fine-grained questions, it enables targeted prompt refinements that explicitly emphasize missing or incorrect components in a generated video. Following previous work in prompt refinements [13, 30], we design an iterative generation loop in which an initial video is generated with Wan 2.2 TI2V-5B [45] and evaluated using PQSG. Based on the PQSG feedback, the prompt is refined by GPT-5.5 and a new video is generated. The resulting video is then re-evaluated using the same PQSG, and the process repeats for multiple iterations. As shown in Fig. 4, from iteration 0 to iteration 1, the average PQSG score increases by nearly 15%, indicating that a single round of targeted prompt refinement substantially improves generation quality.

Table 8: Ablation of PQSG design choices. *Without fine-grained questions* represents the average of three direct VQA scores across objects, actions, and physics.

	Pearson's r	
Metric	w/ GPT-5 QA	w/ Human QA
PQSG (Ours)	0.48	0.80
Without graph dependencies	0.44	0.75
Without fine-grained questions	0.40	0.68

Performance continues to improve in iteration 2, reaching a final average score of 81.9%, after which it plateaus with subsequent refinements. As a baseline, we evaluate the videos Videophy2-AutoEval with the same configuration as Table 1. The Videophy-2-AutoEval baseline shows a similar score improvement across iterations, with small gains throughout. These results demonstrate that PQSG is not only effective as an evaluation framework, but can also be integrated into an iterative refinement loop to improve video generation quality. Crucially, these improvements are achieved through prompt-level augmentations guided by fine-grained PQSG feedback, without modifying the model architecture or retraining, highlighting PQSG's utility as a practical tool for producing higher-quality videos.

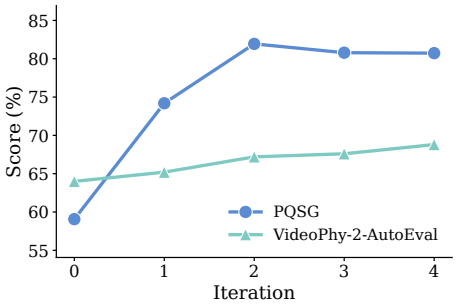

Fig. 4: Scores vs. number of iterations on PQSG refinement loop.

5.6 Ablation Study

We study the design choices of PQSG by changing one component at a time, measuring Pearson's r between the overall score and human Likert score. We experiment with adding a reference video as an additional input during the QG stage, removing the dependency graph (i.e., not automatically marking responses to invalid question as "no"), and removing fine-grained questions (i.e., directly scoring 3 category scores with direct VQA and averaging). As shown in Table 8, each of these ablations degrades performance. Removing the fine-grained questions and removing the dependency graph both reduce correlation with human judgment, confirming that the granular question decomposition and the logical dependency structure each contribute to PQSG's alignment with human ratings.

6 Conclusion

We introduced PQSG, an evaluation framework designed to address a critical gap in video generation: the lack of fine-grained, automated metrics for physical realism. We decompose evaluation into a structured dependency graph of atomic questions, and collect a new dataset FinePhyEval, including fine-grained human annotations of videos from strong video generation models, to facilitate research in developing reliable evaluation metrics. In our experiments, we analyze correlations between human judgments and PQSG scores, compare video generation methods, verify question generation and question answering components, and conduct ablations on design choices.

Limitations. PQSG's performance is bounded by the QA capability of the underlying VLM, where Table 5 demonstrates it currently achieves 64.6% accuracy on physics questions. However, PQSG is model-agnostic by design: as shown in Appendix, human correlation scales with VLM capability, and when humans perform QA, Pearson's correlation reaches 0.80 (Table 8), indicating a high ceiling. Besides, while PQSG improves the reliability of evaluation compared to using direct VQA, the backbone VLM performance is still important. We primarily use closed-source VLMs for our experiments, which may limit reproducibility; we mitigate this by releasing all code, prompts, and annotations, and show generalization with an open-source VLM in Table 7. Finally, PQSG only verifies interactions entailed by the prompt, so physics in regions the prompt leaves unconstrained is not assessed. Evaluating such cases without a prompt, as humans can, is a promising direction for future work.

Acknowledgments

This work was supported by ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, NSF-AI Engage Institute DRL2112635, NSF-CAREER Award 1846185, Capital One Research Award, and NVIDIA Academic Grant Program. The views contained in this article are those of the authors and not of the funding agency.

References

1. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062 (2025)
2. Artificial Analysis: Text to video model arena (2025), https://artificialanalysis.ai/text-to-video/arena?tab=leaderboard-text, accessed: Sept 1, 2025
3. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation (2025)
4. Bao, F., Xiang, C., Yue, G., He, G., Zhu, H., Zheng, K., Zhao, M., Liu, S., Wang, Y., Zhu, J.: Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233 (2024)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. CoRR (2023)
6. Cho, J., Hu, Y., Baldridge, J.M., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In: ICLR (2024)
7. Choi, Y.: Svad: From single image to 3d avatar via synthetic data generation with video diffusion and data augmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3137–3147 (2025)
8. Chow, W., Mao, J., Li, B., Seita, D., Guizilini, V.C., Wang, Y.: Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In: The Thirteenth International Conference on Learning Representations (2025)
9. Fu, X., Wang, X., Liu, X., Bai, J., Xu, R., Wan, P., Zhang, D., Lin, D.: Learning video generation for robotic manipulation with collaborative trajectory control. arXiv preprint arXiv:2506.01943 (2025)
10. Gemini Team: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
11. Gillman, N., Herrmann, C., Freeman, M., Aggarwal, D., Luo, E., Sun, D., Sun, C.: Force prompting: Video generation models can learn and generalize physics-based control signals. arXiv preprint arXiv:2505.19386 (2025)
12. Google: Veo 3 (May 2025), https://aistudio.google.com/models/veo-3
13. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. NeurIPS **36**, 66923–66939 (2023)
14. Hao, Y., Chen, C., Mian, A.S., Xu, C., Liu, D.: Enhancing physical plausibility in video generation by reasoning the implausibility. arXiv preprint arXiv:2509.24702 (2025)
15. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., et al.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation (2024)
16. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: NIPS (2017)

18. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models (2022)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers (2022)
21. Huang, Y., Wang, Z., Lin, H., Kim, D.K., Omidshafiei, S., Yoon, J., Cho, J., Zhang, Y., Bansal, M.: Phymotion: Structured 3d motion reward for physics-grounded human video generation. arXiv preprint arXiv:2605.14269 (2026)
22. Huang, Y., Wang, Z., Lin, H., Kim, D.K., Omidshafiei, S., Yoon, J., Zhang, Y., Bansal, M.: Planning with sketch-guided verification for physics-aware video generation. arXiv preprint arXiv:2511.17450 (2025)
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
24. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and brain sciences **40**, e253 (2017)
25. Lee, D., Yoon, J., Cho, J., Bansal, M.: Self-correcting text-to-video generation with misalignment detection and localized refinement. arXiv preprint arXiv:2411.15115 (2024)
26. Li, C., Michel, O., Pan, X., Liu, S., Roberts, M., Xie, S.: Pisa experiments: Exploring physics post-training for video diffusion models by watching stuff drop. arXiv preprint arXiv:2503.09595 (2025)
27. Lin, W., Jia, L., Hu, W., Pan, K., Yue, Z., Zhao, W., Chen, J., Wu, F., Zhang, H.: Reasoning physical video generation with diffusion timestep tokens via reinforcement learning. arXiv preprint arXiv:2504.15932 (2025)
28. Liu, S., Ren, Z., Gupta, S., Wang, S.: Physgen: Rigid-body physics-grounded image-to-video generation. In: European Conference on Computer Vision (ECCV) (2024)
29. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models (2023)
30. Mañas, O., Astolfi, P., Hall, M., Ross, C., Urbanek, J., Williams, A., Agrawal, A., Romero-Soriano, A., Drozdzal, M.: Improving text-to-image consistency via automatic prompt optimization. Transactions on Machine Learning Research (2024), `https://openreview.net/forum?id=g12Gdl6aDL`, featured Certification
31. Meng, F., Liao, J., Tan, X., Lu, Q., Shao, W., Zhang, K., Cheng, Y., Li, D., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In: ICML (2025)
32. Montanaro, A., Savant Aira, L., Aiello, E., Valsesia, D., Magli, E.: Motioncraft: Physics-based zero-shot video generation. Advances in Neural Information Processing Systems **37**, 123155–123181 (2024)
33. Motamed, S., Culp, L., Swersky, K., Jaini, P., Geirhos, R.: Do generative video models understand physical principles? (2025)
34. Ni, H., Egger, B., Lohit, S., Cherian, A., Wang, Y., Koike-Akino, T., Huang, S.X., Marks, T.K.: Ti2v-zero: Zero-shot image conditioning for text-to-video diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9015–9025 (2024)

35. OpenAI: Video generation models as world simulators (2024), https://openai.com/index/video-generation-models-as-world-simulators/
36. OpenAI: GPT-5 model (2025), https://openai.com/gpt-5/
37. OpenAI: Sora 2 (October 2025), https://openai.com/index/sora-2/
38. Qin, Y., Shi, Z., Yu, J., Wang, X., Zhou, E., Li, L., Yin, Z., Liu, X., Sheng, L., Shao, J., et al.: Worldsimbench: Towards video generation models as world simulators. In: Forty-second International Conference on Machine Learning (2025)
39. Ross, C., Hall, M., Romero-Soriano, A., Williams, A.: What makes a good metric? evaluating automatic metrics for text-to-image consistency. In: First Conference on Language Modeling (2024), https://openreview.net/forum?id=LFfktMPAci
40. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans (2016), https://arxiv.org/abs/1606.03498
41. Soni, A., Venkataraman, S., Chandra, A., Fischmeister, S., Liang, P., Dai, B., Yang, S.: Videoagent: Self-improving video generation for embodied planning. In: Workshop on Reinforcement Learning Beyond Rewards@ Reinforcement Learning Conference 2025 (2025)
42. Spelke, E.S.: Principles of object perception. Cognitive science **14**(1), 29–56 (1990)
43. Spelke, E.S., Gutheil, G., Van de Walle, G.: The development of object perception. Visual Cognition (1995)
44. Tan, X., Jiang, Y., Li, X., Zong, Z., Xie, T., Yang, Y., Jiang, C.: Physmotion: Physics-grounded dynamics from a single image. arXiv preprint arXiv:2411.17189 (2024)
45. Team Wan: Wan: Open and advanced large-scale video generative models (2025)
46. Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., Neubig, G.: Do llms exhibit human-like response biases? a case study in survey design. Transactions of the Association for Computational Linguistics **12**, 1011–1026 (2024)
47. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards Accurate Generative Models of Video: A New Metric & Challenges (2019), https://arxiv.org/abs/1812.01717
48. Wang, C., Chen, C., Huang, Y., Dou, Z., Liu, Y., Gu, J., Liu, L.: Physctrl: Generative physics for controllable and physics-grounded video generation. arXiv preprint arXiv:2509.20358 (2025)
49. Wang, P., Wang, W., Li, Q.: Physcorr: Dual-reward dpo for physics-constrained text-to-video generation with automated preference selection. arXiv preprint arXiv:2511.03997 (2025)
50. Wang, Z., Cho, J., Li, J., Lin, H., Yoon, J., Zhang, Y., Bansal, M.: Epic: Efficient video camera control learning with precise anchor-video guidance. arXiv preprint arXiv:2505.21876 (2025)
51. Wang, Z., Lin, H., Yoon, J., Cho, J., Zhang, Y., Bansal, M.: Anchorweave: World-consistent video generation with retrieved local spatial memories. arXiv preprint arXiv:2602.14941 (2026)
52. Yang, X., Li, B., Zhang, Y., Yin, Z., Bai, L., Ma, L., Wang, Z., Cai, J., Wong, T.T., Lu, H., et al.: Vlipp: Towards physically plausible video generation with vision and language informed physical prior. arXiv preprint arXiv:2503.23368 (2025)
53. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: The Thirteenth International Conference on Learning Representations (2024)

54. Yin, Z., Chen, K., Bai, X., Jiang, R., Li, J., Li, H., Liu, J., Xiang, Y., Yu, J., Zhang, M.: A survey: spatiotemporal consistency in video generation. ACM Computing Surveys **58**(12), 1–41 (2026)
55. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)