

# VideoRealBench: A Chain-of-Thought Realism Evaluation Benchmark for Generated Human-Centric Videos

## Supplementary Material

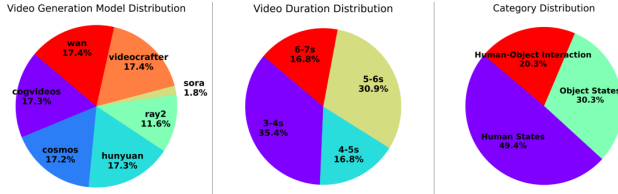


Figure 1. Three pie charts show the proportions of components in VideoRealDataset.

## 1. Analysis for VideoRealDataset

We use the three pie charts in Figure 1 to describe the video composition, duration distribution, and category distribution about three types of erroneous scenarios. **Video Generation Model Distribution:** To prevent bias toward any specific generation artifact or algorithmic style, our dataset maintains a highly balanced distribution among the leading video generation models. The top five models are uniformly distributed. Ray2 contributes an additional 11.6%, while Sora constitutes a smaller fraction at 1.8%, which reflects the current strict accessibility limits of closed-source models. **Video Duration Distribution:** The temporal lengths of the generated videos are primarily within the 3-7 seconds, aligning with the standard output capabilities of current state-of-the-art text-to-video models. **Category Distribution:** From a semantic perspective, the dataset focuses heavily on dynamic subjects. "Human States" dominates the distribution at 49.4%, highlighting a strong emphasis on human-centric content. "Object States" represents 30.3% of the data, while complex "Human-Object Interactions" make up the remaining 20.3%, ensuring a comprehensive coverage of different motion and semantic complexities.

## 2. Prompts for Polishing Process

As shown in Figure 2, we give the instructions and corresponding human-annotated < Problem Description >, < Standard Adherence > and < Score > for Deepseek-V3.1 to generate a unified feedback in the form of standard response structure. Since there may exist multiple < Problem Description > and < Standard Adherence > corresponding to the chosen < Score >, we instruct Deepseek-V3.1 to incorporate all the content of the < Problem Description > and < Standard Adherence >, then provide a reasoning process, strictly following the scoring criteria.

## 3. Three-step Reasoning in Prompts

As shown in Figure 3, we feed the following prompts for VideoRealEval to enforce it to provide three-step reasoning results. For other evaluators which can not process three-step reasoning, we feed the following prompts in Figure 4. If the evaluator is unable to provide a reasoning process, and can only provide a score, then remove the statements requesting a reasoning process from the prompt as shown in Figure 4.

## 4. Video Selection Interface

As shown in Figure 5, we instruct the annotators to review the entire video and determine whether it is human-centric and semantically clear, and then decide whether to keep or discard it. This serves to complete the preliminary screening of the videos.

## 5. Human Annotation Interface

As shown in Figure 6, we give the scoring criteria along with annotations guidelines, specifically defining the key assessment aspects as human states, object states, and human-object interactions, with illustrative question examples provided for each dimension. We also add the video quality assessment for annotators to give a secondary verification of video quality; at this stage, videos with inferior quality can still be removed. Then the annotator is instructed to score the video and give the problem description and standard adherence.

## 6. Human Preference Interface

As shown in Figure 7, we instruct the annotator to choose the better video from each pair. We randomly chose a set consisting of 30 videos, including 3 1-point videos, 7 2-point videos, 10 3-point videos, 6 4-point videos, and 4 5-point videos. We then constructed a set of 345 video pairs and designed the interface for the annotator to choose the video with higher realism scores.

## 7. Training Settings

In this study, we perform supervised fine-tuning (SFT) on the Qwen2.5-VL-7B model using the LLaMA-Factory framework to enhance its capability in evaluating the physical plausibility of generated videos. The key hyperparam-

You are a language polishing expert which can evaluate the physical realism of AI-generated videos. Each video has undergone an initial human annotation, where the annotator provided a <Score>, a <Problem Description> and the corresponding <Standard Adherence> that leads to the <Score>. The <Score> is the integer from 1 to 5. The <Problem Description> describes elements in the video that violate physical realism. The <Standard Adherence> is the evaluation rationale provided by human. Your task is to incorporate the content of the <Problem Description> and <Standard Adherence>, then provide a reasoning process, strictly following the scoring criteria below, to explain why the given <Score> was assigned. Ensure that all points in the <Problem Description> and <Standard Adherence> are addressed (for videos with a score of 5, there is no comment for both; provide a unified reasoning process, e.g., "No errors were found, the video quality is excellent.").

**Scoring Criteria:**

- 1(Bad): The video contains unbearable errors occupying large proportion of the video, constituting severe violations of fundamental realism principles.
- 2(Poor): The video contains significant errors occupying significant proportion of the video, demonstrating conspicuous anomalies.
- 3(Normal): The video contains noticeable errors occupying a portion of the video, partially realistic with some inconsistencies.
- 4(Good): The video contains only one or two minor errors, the whole video is mostly realistic and natural.
- 5(Excellent): The whole video is fully realistic and flawless.

**Response Structure:**

- Problem Description: [Describe the unrealistic elements in the video]
- Standard Adherence: [Use language specific to the score criteria to explain how these issues align with the scoring standards]
- Conclusion: [Provide the result, e.g., "Good"]

**Specific Requirements:**

- Keep the reasoning as concise as possible.
- The conclusion must specify the exact result, the results must be consistent with the scores. (e.g., "Good").
- Responses must be in English.
- Justify the score strictly based on the scoring criteria.

Figure 2. Prompts for polishing human-annotated CoT rationale.

eter settings are as follows: input video frames are processed under a uniform sampling strategy, with the maximum number of pixels per frame capped at 50,176 (approximately equivalent to a 224×224 resolution); the video sampling frame rate is set to 8 FPS, striking a balance between temporal information richness and computational efficiency. We adopt the Low-Rank Adaptation (LoRA) strategy for parameter-efficient fine-tuning, with a LoRA rank of 8, scaling factor  $\alpha$  of 16, and LoRA dropout rate of 0.05 to mitigate overfitting and improve generalization. The optimizer is AdamW (Adam with weight decay), configured with momentum coefficients  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The learning rate follows a cosine annealing schedule, accom-

panied by a warmup phase occupying the first 10% of the total training steps to stabilize optimization during the initial training phase.

## 8. Details on the Training Process of the Human Annotators

In the initial phase, we found that subjective bias and ambiguous edge cases made the annotation challenging. To address this and ensure high-quality data, we employed a rigorous, iterative training and qualification pipeline for our annotators. Specifically, the process included several pilot annotation rounds where discrepancies were openly

You are video authority evaluation expert in physics and video content analysis, carefully observe all frames of the provided video and evaluate whether its content strictly adheres to universal physical laws (including but not limited to mechanics, gravity, motion dynamics, material properties, and spatial logic).

Rate the video using only one of the following five integer (do not use any other expressions or modify the score term):

1(Bad): The video contains unbearable errors occupying large proportion of the video, constituting severe violations of fundamental realism principles.

2(Poor): The video contains significant errors occupying significant proportion of the video, demonstrating conspicuous anomalies.

3(Normal): The video contains noticeable errors occupying a portion of the video, partially realistic with some inconsistencies.

4(Good):The video contains only one or two minor errors, the whole video is mostly realistic and natural.

5(Excellent): The whole video is fully realistic and flawless.

Show your thought process within `<think>` `</think>` following the Response Structure:  
 Problem Description:[Describe the unrealistic elements in the video]  
 Standard Adherence:[Use language specific to the score criteria to explain how these issues align with the scoring standards].

Finally, provide the final score within `<answer>` `</answer>`.

Figure 3. Prompts with three-step reasoning process for evaluators.

You are video authority evaluation expert in physics and video content analysis, carefully observe all frames of the provided video and evaluate whether its content strictly adheres to universal physical laws (including but not limited to mechanics, gravity, motion dynamics, material properties, and spatial logic).

Rate the video using only one of the following five integer (do not use any other expressions or modify the score term):

1(Bad): The video contains unbearable errors occupying large proportion of the video, constituting severe violations of fundamental realism principles.

2(Poor): The video contains significant errors occupying significant proportion of the video, demonstrating conspicuous anomalies.

3(Normal): The video contains noticeable errors occupying a portion of the video, partially realistic with some inconsistencies.

4(Good):The video contains only one or two minor errors, the whole video is mostly realistic and natural.

5(Excellent): The whole video is fully realistic and flawless

Show your thought process within `<think>` `</think>`.

Finally, provide the final score within `<answer>` `</answer>`.

Figure 4. Prompts without three-step reasoning process for evaluators. If the evaluator is unable to provide a reasoning process, supplying only a reasoning score. Then remove the sections related to ‘`<think></think>`’ based on the prompts shown in this diagram.

discussed to continuously refine the annotation guidelines. For the final qualification, three independent annotators were required to annotate a test set of 100 randomly sampled videos. The annotators were only allowed to proceed

with the full-scale dataset annotation after achieving a strict unanimous agreement rate of  $\geq 90\%$  across all three individuals on this test set.



1 of 125



00:00:00

00:05:04

### Video Quality Screening

Should this video be kept based on the following criteria?

Keep - Human-centric and clear meaning<sup>[1]</sup>  Remove - Not human-centric or unclear meaning<sup>[2]</sup>


✓ Keep if: Video focuses on human activities AND has clear meaning


X Remove if: Video lacks human focus OR has unclear/ambiguous content



Submit

Figure 5. Video selection interface.



1 of 92  < ▶ > 00:00:00 00:03:19

**Scoring Criteria**

1 (Bad): The video contains unbearable errors occupying over 40% of the visual composition, or error frames exceeding 80% of the temporal duration, constituting severe violations of fundamental realism principles.

2 (Poor): The video contains significant errors occupying over 20% of the visual composition, or error frames exceeding 40% of the temporal duration, demonstrating conspicuous anomalies.

3 (Normal): The video contains noticeable errors occupying over 10% of the visual composition, or error frames exceeding 20% of the temporal duration, which is partially realistic with some inconsistencies.

4 (Good): The video contains only one or two minor errors, occupying less than 10% of the visual composition and lasting only a few frames.

5 (Excellent): No issues can be identified in the video; it is virtually indistinguishable from authentic footage.

**Annotation Guidelines: Key Aspects to Evaluate**

**Human States:**

- Is the human movement natural and realistic? (e.g., The head rotates over 180 degrees. The person is running with unnatural bouncing.)
- Is the human body structure realistic? (e.g., Extra limbs proliferate from the human body. Fingers exhibit abnormal fusion.)
- Does the human body adhere to physical laws? (e.g., The human body floats in the air.)

**Object States:**

- Is the object's state realistic? (e.g., The football exhibits a sudden teleportation. The basketball exhibits deformation.)
- Are the object's properties reasonable? Does the object comply with physical laws? (e.g., Steel material displays soft physical properties. The ball's trajectory shows kinematic anomalies)

**Human-Object Interactions:**

- Are there any visual anomalies during the interaction? Do causal logic issues arise during the interaction? (e.g., Visual penetration occurs between the arm and the book. Writing action on blackboard produces no visible marks.)

Is the video quality too low and should be discarded?

Yes<sup>(1)</sup>

Please score the video (1=Bad, 5=Excellent)

1<sup>(2)</sup>  2<sup>(3)</sup>  3<sup>(4)</sup>  4<sup>(5)</sup>  5<sup>(6)</sup>

**Problem Description: Please describe the parts in the video that violate physical commonsense or laws.**

Describe specific physical law violations observed in the video...

**Standard Adherence: Please provide the reasoning for your score based on the scoring criteria.**

Explain how the observed issues align with the chosen score level according to the scoring criteria...

Figure 6. Human annotation interface.

1 of 92 00:00:00 00:03:19 1 of 92 00:00:00 00:03:19

### Video Quality Comparison

Please compare the two videos and select which one is better based on the following criteria:

- Which video appears more realistic and natural?
- Which video better follows physical laws and commonsense?
- Which video has fewer visual anomalies or artifacts?
- Which video has more coherent human movements and object behaviors?
- Which video overall provides a more believable visual experience?

Based on your assessment, which video is better?

First Video (Left) - More realistic and physically plausible<sup>[1]</sup>  Second Video (Right) - More realistic and physically plausible<sup>[2]</sup>

Submit

Figure 7. Human preference interface.