Video-Bench: Human-Aligned Video Generation Benchmark

Supplementary Material

In Sec. A, we elaborate on our evaluation dimensions with comprehensive explanations and examples. Sec. B presents extended experimental results and analysis. Sec. C discusses the broader implications and potential impacts on society. Finally, Sec. D examines the current limitations and outlines promising directions for future research.

A. More Details on Evaluation Dimension

A.1. Video-Condition Consistency

A.1.1. Object Class Consistency

Definition and scope Object class consistency evaluation assesses ***the consistency between objects in the video and those specified in the text prompt***. The assessment should consider the following key aspects:

- **Generation accuracy**: Whether objects mentioned in the text are correctly generated.
- **Class identification**: Whether object categories are clearly identifiable.
- **Appearance fidelity**: Whether generated objects' appearance and structure align with objective reality and human perceptual expectations.
- **Deformation**: Whether objects maintain their structural integrity during motion.

Scoring criteria

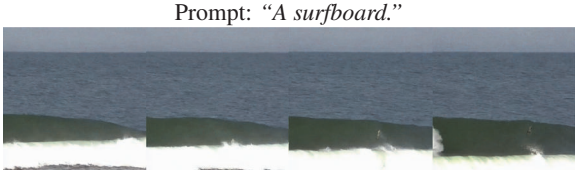
- ❶ **Poor consistency**: Objects are completely unrecognizable or fail to match the specified objects in the prompt.
- ❷ **Moderate consistency**: Objects are barely recognizable as the specified class but exhibit one or more of the following issues:
  - Partial object generation (*e.g.*, only a hand visible when a complete person is specified).
  - Feature mixing between specified object and other object classes.
  - Unstable object characteristics (*e.g.*, facial features appearing and disappearing).
  - Presence of unspecified objects of the same category or multiple similar objects occupying significant space (*e.g.*, a motorcycle consistently appearing alongside a specified car).
- ❸ **Good consistency**: Object classes remain correct and consistent throughout the entire video, avoiding all issues mentioned in the moderate consistency category.

A.1.2. Action Consistency

Definition and scope Action consistency evaluation assesses ***the consistency between actions in the video and those specified in the text prompt***. The assessment should consider the following key aspects:

- **Generation accuracy**: Whether actions mentioned in the text are correctly generated.
- **Action identification**: Whether actions are clearly identifiable.
- **Process fidelity**: Whether the appearance and progression of actions align with objective reality and human perceptual expectations.

Prompt: *"A surfboard."*



(a) Poor object class consistency (Score=1)

(b) Moderate object class consistency (Score=2)

(c) Good object class consistency (Score=3)

Figure 4. **Comparative examples of object class consistency assessment. (a) Poor**: Generated scene shows only ocean waves without any surfboard, completely failing to meet the prompt requirement. **(b) Moderate**: Multiple surfboards present but with additional decorative elements and patterns that complicate the straightforward prompt requirement. **(c) Good**: Clean white surfboard rendered consistently throughout the sequence, precisely matching the simple prompt specification.

Scoring criteria

- ❶ **Poor consistency**: Actions are either completely unrecognizable or incorrectly generated.
- ❷ **Moderate consistency**: Actions are partially consistent but exhibit one or more of the following issues:
  - Significant deviation from the realistic appearance or progression of the action.
  - Incomplete action representation, either in terms of viewpoint or temporal coverage, showing only a fragment of the complete action.
- ❸ **Good consistency**: Actions fully align with the prompt specifications and avoid all issues mentioned in the moderate consistency category.

Important notes

- This metric focuses primarily on the presence and accuracy of actions in the video rather than their dynamic presentation or motion effects. However, completely static videos that fail to show any movement should be scored as inconsistent with objective understanding.

A.1.3. Color Consistency

Definition and scope Color consistency evaluation assesses ***the degree of color matching between the video and the provided text prompt***. The assessment should consider the following key aspects:

- **Color consistency**: Whether colors align with the text prompt and maintain stability throughout the video with-

Prompt: "*A person is marching.*"

(a) Poor action consistency (Score=1)

(b) Moderate action consistency (Score=2)

(c) Good action consistency (Score=3)

Figure 5. **Comparative examples of action consistency assessment. (a) Poor**: The man appears to be walking naturally in the video rather than marching, failing to demonstrate any marching motion, contradicting the prompted action completely. **(b) Moderate**: The figure shows walking/sliding motion, but the movement appears unnatural and doesn't fully capture the rhythmic, structured nature of marching. **(c) Good**: Clear marching action with proper leg movement and posture, displaying fluid and natural progression of steps that precisely matches the prompted action.

Prompt: "*A yellow cat.*"

(a) Poor color consistency (Score=1)

(b) Moderate color consistency (Score=2)

(c) Good color consistency (Score=3)

Figure 6. **Comparative examples of color consistency assessment. (a) Poor**: The generated cat appears brown/orange with unnatural blue-green eyes, significantly deviating from the prompted yellow color. **(b) Moderate**: Cat displays an orange/ginger coloration that, while consistent throughout the sequence, falls within a similar but distinct color spectrum from the requested yellow. **(c) Good**: The cat exhibits pure yellow coloring that precisely matches the prompt specification, maintaining a consistent hue throughout the sequence.

out sudden changes.
- **Color placement**: Whether colors appear on the correct objects or within appropriate scenes.

Scoring criteria

- ❶ **Poor consistency**: Objects are either incorrectly generated or display colors that completely deviate from the text prompt specifications.
- ❷ **Moderate consistency**: Correct colors appear in the video but exhibit imperfections in one or more of the following aspects:
  - Incorrect color allocation (*e.g.*, colors appearing in background instead of on intended objects).
  - Color instability with sudden changes or variations in object coloring.
  - Color confusion where objects display correct colors mixed with significant areas of unintended colors (*e.g.*, a requested white vase generated as black and white).
  - Poor color distinction between objects and background.
  - Approximate color matching within the same spectrum but lacking precision (*e.g.*, pink versus purple, yellow versus orange).
- ❸ **Good consistency**: Colors demonstrate high fidelity to the text prompt, maintain stability throughout the video, show correct distribution, and exhibit no sudden changes or inconsistencies. The work avoids all issues mentioned in the moderate consistency category.

A.1.4. Scene Consistency

Definition and scope Scene Consistency evaluation assesses ***the consistency between scenes in the video and those specified in the text prompt***. The assessment should consider the following key aspects:

- **Generation accuracy**: Whether scenes mentioned in the text are correctly generated.
- **Scene identification**: Whether scenes are clearly identifiable.
- **Element fidelity**: Whether the appearance and structure of scene elements align with objective reality and human perceptual expectations.

Scoring criteria

- ❶ **Poor consistency**: Scene generation is completely unrelated to the text prompt and scenes are difficult to identify.
- ❷ **Moderate consistency**: Scenes are barely recognizable and exhibit one or more of the following issues:
  - Partial scene generation without showing the complete scene context.
  - Display of limited scene characteristics (*e.g.*, only bread in a bakery, only a sink in a bathroom).
  - Scene generation that is similar but not precisely matching the specified scene.
- ❸ **Good consistency**: Scenes are clearly identifiable and align with human subjective understanding of objective world arrangements.

Important notes

- For ambiguous scene terms, scoring should use the most comprehensive interpretation among the generated results as the standard. For example, if "bathroom" is generated as a complete bathroom with a bathtub by one model and as a simple washroom with only a mirror, sink, or toilet

by another, the complete bathroom setting should be used as the reference standard.

Prompt: "*Arch.*"

(a) Poor scene consistency (Score=1)

(b) Moderate scene consistency (Score=2)

(c) Good scene consistency (Score=3)

Figure 7. **Comparative examples of scene consistency assessment. (a) Poor**: The scene displays an architectural structure but lacks clear arch definition, with distracting foreground elements and inconsistent structural representation. **(b) Moderate**: The sunset scene through an arch frame demonstrates a recognizable arch structure, though the dramatic lighting and silhouette effect partially obscure architectural details. **(c) Good**: Garden arch with natural foliage shows a clear, well-defined arch structure maintained consistently throughout the sequence, with proper architectural form and depth.

A.1.5. Video-text Consistency

Definition and scope Video-text consistency evaluates ***the comprehensive alignment between the video and the text prompt***. The assessment should consider the following key aspects:

- **Core element coverage**: Whether the video demonstrates all core elements mentioned in the text prompt (including humans, animals, actions, objects, scenes, styles, spatial relationships, and quantitative relationships).
- **Visual clarity**: Whether the video's image quality affects comprehension of its content.

Scoring criteria

- ❶ **Very poor consistency**: Missing half or more of the key elements, demonstrating very weak consistency, or visual quality so poor that video comprehension is impossible.
- ❷ **Poor consistency**: Video includes most key elements but most are insufficiently generated, or visual quality is inadequate for determining consistency with the text prompt.
- ❸ **Moderate consistency**: Video either includes most key elements with sufficient generation or includes all elements but most are insufficiently generated. Visual quality is adequate for determining consistency with the text prompt.
- ❹ **Good consistency**: Video includes all key elements, with some elements insufficiently generated. Visual quality is adequate for determining consistency with the text prompt.
- ❺ **Excellent consistency**: Video includes all key elements with sufficient generation and complete alignment with the text prompt. Visual quality is adequate for determining consistency with the text prompt.

Important notes

- Insufficient generation refers to elements that are present but fail to meet consistency requirements, such as low visibility in actions or objects that don't conform to objective world appearances.
- "Most" is determined by the number of key elements in the prompt, typically not exceeding 5 elements, thus "most" generally means $N - 1$ elements.
- This metric does not have high requirements for visual quality. Superior visual quality is not a prerequisite for high scores.

Prompt: "*Two pandas discussing an academic paper.*"

(a) Very poor video-text consistency (Score=2)

(b) Moderate video-text consistency (Score=3)

(c) Excellent video-text consistency (Score=5)

Figure 8. **Comparative examples of video-text consistency assessment. (a) Very poor**: Shows only a single panda with no interaction or academic elements, significantly deviating from the prompt requirements. **(b) Moderate**: While two pandas are shown and appear to be facing each other, the interaction lacks clear indication of academic discussion or the presence of a paper. **(c) Excellent**: The scene perfectly captures the prompt, showing two pandas in an interactive pose with what appears to be a paper between them, complete with a natural academic discussion setting.

A.2. Video Quality

A.2.1. Imaging Quality

Imaging quality evaluates ***the visual fidelity and clarity of the generated video compared to standard high-definition content***. The assessment should consider the following key aspects:

- **Image clarity**: Overall sharpness, resolution, and detail preservation throughout the video.
- **Visual artifacts**: Presence of noise, distortion, overexposure, or other technical imperfections.

Scoring criteria

- ❶ **Very poor quality**: Severe visual artifacts with obvious distortions, extreme blurriness, excessive noise, and significant overexposure issues that severely impact the viewing experience.
- ❷ **Poor quality**: Notable visual artifacts with apparent distortions, general blurriness, and noise that detract from the natural appearance and viewing experience.
- ❸ **Moderate quality**: Resolution comparable to 480p standard definition, with minor artifacts, slight noise, and occasional exposure issues that moderately affect the viewing experience.
- ❹ **Good quality**: Resolution comparable to 720p high definition, with minimal artifacts and generally pleasant viewing experience.
- ❺ **Excellent quality**: Resolution comparable to 1080p full HD or better, with no discernible artifacts, providing an exceptional viewing experience comparable to professional video content.

Prompt: "*A person is jogging.*"

(a) Very poor imaging quality (Score=1)

(b) Excellent imaging quality (Score=5)

Figure 9. **Comparative examples of imaging quality assessment. (a) Very poor**: The generated video exhibits poor video clarity with visible noise, unstable camera movement, and inconsistent lighting conditions. **(b) Excellent**: It demonstrates superior visual fidelity with stable framing, natural lighting, sharp details, and professional cinematographic quality.

A.2.2. Aesthetic Quality

Definition and scope Aesthetic quality evaluation encompasses ***the artistic and compositional elements of video production***, including structural arrangement, color utilization, compositional effectiveness, visual appeal, and overall harmonic integration. The assessment of video-text consistency should consider the following key aspects:

- **Structural coherence**: Whether the arrangement and composition of subjects (people or objects) in the video are logically sound and aesthetically pleasing, rather than causing psychological discomfort.
- **Color application**: The appropriateness and effectiveness of color usage throughout the video sequence.
- **Compositional efficacy**: Whether the composition effectively captures and presents all necessary information specified in the text prompt.
- **Visual appeal**: The video's capacity to maintain visual engagement.
- **Overall harmony**: The degree to which all elements work together cohesively to create a unified and harmonious visual experience.

Scoring criteria

- ❶ **Very poor aesthetic quality**: The work exhibits severe deficiencies in color utilization, composition, and clarity. It lacks visual appeal and emotional expression, with poor overall harmonic integration.
- ❷ **Poor aesthetic quality**: The work demonstrates notable deficiencies in specific aspects, such as discordant color schemes or inadequate composition, significantly compromising the overall aesthetic experience.
- ❸ **Moderate aesthetic quality**: The work shows average performance across most dimensions, with possible minor deficiencies in certain aspects, while maintaining a basic aesthetic experience.
- ❹ **Good aesthetic quality**: The work demonstrates strong execution in color usage, composition, and clarity, delivering a satisfying visual experience with appropriate emotional expression and creative elements.
- ❺ **Excellent aesthetic quality**: The work excels in all aspects, achieving high standards in color utilization, composition, and clarity. It delivers powerful visual impact and profound emotional expression, providing an outstanding aesthetic experience.

Prompt: "*A bear is climbing trees.*"

(a) Very poor aesthetic quality (Score=1)

(b) Excellent aesthetic quality (Score=5)

Figure 10. **Comparative examples of aesthetic quality assessment. (a) Very poor**: Bear climbing tree sequence with flat composition, lacking visual depth and artistic consideration in framing and lighting. **(b) Excellent**: Cat leaping sequence captured in golden-hour lighting with atmospheric forest backdrop, demonstrating sophisticated composition and cinematic appeal.

A.2.3. Temporal Consistency

Definition and scope Temporal consistency evaluates ***the consistency of semantic and visual features between consecutive frames*** in the video, ensuring smooth transitions without abrupt changes or unnatural jumps. The assessment encompasses two primary aspects:

- **Visual feature consistency**:
  - **Color and brightness**: Smooth transitions between consecutive frames without flickering or sudden changes in illumination.
  - **Texture and detail**: Maintenance of consistent object textures and details across frames without unexpected blur or clarity shifts.
- **Semantic consistency**:
  - **Object position and shape**: Preservation of consistent object positioning and morphology between frames without unnatural deformation or displacement.
  - **Scene layout**: Maintenance of consistent scene composition and background elements across frames.
  - **Subject coherence**: Stability of main subjects across consecutive frames without abrupt changes or unnatu-

ral transitions.

Scoring criteria

- ❶ **Very poor consistency**: Significant inconsistencies in color, brightness, and texture between frames with obvious flickering or sudden changes. Semantic features show discrepancies in object positioning and scene layout, with main subjects exhibiting sudden or unnatural variations.
- ❷ **Poor consistency**: Notable inconsistencies in visual features. Semantic features maintain general consistency but occasionally display issues with object positioning and scene layout. Main subjects may exhibit minor inconsistencies.
- ❸ **Moderate consistency**: Visual features typically maintain consistency with minor fluctuations in color, brightness, and texture. Semantic features show general consistency with slight issues affecting object position, shape, and scene layout coherence. Main subjects maintain general consistency with minor deviations.
- ❹ **Good consistency**: Visual features maintain consistency between frames with smooth transitions in color, brightness, and texture. Semantic features demonstrate coherence with stable object positions, shapes, and scene layout. Main subjects maintain consistency with only minor inconsistencies that don't significantly impact overall coherence.
- ❺ **Excellent consistency**: All visual features demonstrate seamless consistency between frames without perceptible flickering or sudden changes. Semantic features show complete consistency in object positions, shapes, scene layout, and background. Main subjects maintain perfect consistency without notable deviations that would affect viewer perception of continuity.

Prompt: "*A person is skateboarding.*"

(a) Very poor temporal consistency (Score=1)

(b) Excellent temporal consistency (Score=5)

Figure 11. **Comparative examples of temporal consistency assessment. (a) Very poor:** Skateboarding sequence shows inconsistent object placement and scaling between frames, with abrupt position shifts disrupting motion continuity. **(b) Excellent:** Skateboarding motion displays coherent object positioning and consistent spatial relationships across frames, maintaining seamless temporal flow.

A.2.4. Motion Effects

Definition and scope Motion effects evaluate ***the quality of subject motion and its interaction with the environment*** in the video. The assessment should consider the following key aspects:

- **Physical accuracy**: Whether object motion trajectories conform to physical laws such as inertia and gravity.
- **Dynamic blur**: Whether motion blur appropriately corresponds to the speed and direction of movement.
- **Environmental interaction**: Whether the relationship between moving objects and their background is coherent, including expected occlusion and reflections.
- **Lighting physics**: Whether changes in shadows and lighting during object movement align with physical laws, enhancing scene realism.

Scoring criteria

- ❶ **Very poor effects**: Motion trajectories are severely incorrect, or the primary characteristics of movement are generated so poorly that the motion is barely recognizable. Clear violations of physical laws are present, and dynamic blur is either absent or completely misaligned with the motion.
- ❷ **Poor effects**: Motion trajectories are poorly generated and movement is barely recognizable. Dynamic blur is inconsistent with movement speed and direction, and there are obvious issues with object interaction with background and lighting.
- ❸ **Moderate effects**: Motion effects are generally present and movement is recognizable, but exhibits one of the following issues:
  - Compromised motion smoothness, with noticeable frame-to-frame inconsistencies or abrupt changes disrupting motion fluidity.
  - Inadequate or excessive dynamic blur application, failing to accurately reflect movement speed and direction.
  - Partially maintained motion consistency, but certain elements like object-environment interaction or lighting changes lack convincing portrayal.
- ❹ **Good effects**: Movement is recognizable, motion trajectories and dynamic blur are mostly coherent, but certain aspects of motion appear unnatural and do not align with human subjective understanding of objective world changes.
- ❺ **Excellent effects**: Movement is clearly recognizable, motion trajectories are accurate, dynamic blur is appropriately applied, and interaction of moving objects with their environment, including shadows and lighting, is seamlessly integrated and realistic.

Important notes

- This metric focuses specifically on dynamic presentation and effects, not on motion consistency with the text prompt. Consistency does not affect the scoring.
- In this metric, videos displaying static or no motion should be scored as 1 or 2.

B. Additional Experimental Results

B.1. Statistical Analysis of Evaluation Discrepancy

Table 9 presents a comprehensive statistical analysis of the evaluation discrepancies between our MLLM-based framework and human assessments across nine critical dimensions. The results reveal several noteworthy patterns:

- ***Video Quality Metrics:*** Among the four video quality metrics, temporal consistency shows the highest positive mean difference (0.31), suggesting our framework maintains stricter standards in assessing temporal coherence. Interestingly, motion effects exhibit the only negative mean difference (-0.26), indicating human evaluators are more sensitive to motion-related qualities. This

Prompt: "*A person is catching or throwing baseball.*"

(a) Very poor motion Effects (Score=1)

(b) Excellent motion Effects (Score=5)

Figure 12. **Comparative examples of motion effects assessment. (a) Very poor**: Pitcher motion sequence exhibits abrupt transitions with insufficient temporal continuity, resulting in unrealistic motion blur and motion artifacts. **(b) Excellent**: Baseball player's pitching action rendered with smooth frame transitions and natural motion progression, maintaining consistent motion quality throughout the sequence.

contrast highlights the complementary nature of machine and human evaluation capabilities.

- ***Alignment Metrics:*** In the video-condition alignment category, video-text consistency (0.19) and action consistency (0.22) demonstrate the most substantial positive differences. This aligns with our findings in the main paper, where the framework showed superior performance in detecting subtle semantic misalignments. The narrower confidence intervals in object-class consistency $[0.03, 0.05]$ and color consistency $[0.04, 0.06]$ suggest more stable and reliable evaluations in these aspects.
- ***Statistical Significance:*** All confidence intervals at the 99% level exclude zero, indicating statistically significant differences between MLLM and human evaluations across all dimensions. The tight confidence intervals, particularly in object-class and color consistency evaluations, demonstrate the robustness and reliability of our framework. These findings quantitatively support our framework's capability to provide more stringent and consistent evaluations in most dimensions, while also revealing areas where human perception remains uniquely valuable (*e.g.*, motion effects).

B.2. Mini-split for quick performance evaluation

To quickly evaluate video generation performance, we proposed the min-split scheme and conducted a comprehensive test on Sora 8. The results show that Sora excels in video quality, motion effects, and video-condition alignment, outperforming Gen-3 and CogVideoX. However, there are some limitations in video-text consistency, temporal coherence, and motion generation, such as content mismatch, unnatural frame transitions, and inconsistent motion trajectories.

Specifically, we randomly selected 25 representative prompts (about one-third of the dataset) and generated 25 videos with Sora. The generation parameters were set to 720p resolution, 16:9 aspect ratio (to avoid wide-angle distortions from a 1:1 aspect ratio) and 5-second duration. Among the 9 evaluation metrics, Sora ranked first in 5 and demonstrated strong competitiveness across the remaining metrics, showcasing its potential as one of the most advanced video generation models available.

As shown in the results, Sora ranked first in 5 out of the 9 evaluated metrics and performed competitively across the remaining metrics. Notably, in all Video Quality metrics and in Video-Text Consistency, Sora matched or outperformed Gen-3 and CogVideoX, which had previously led these categories by a significant margin. Furthermore, Sora exhibited superior performance in Motion Effects and most Video-Condition Alignment metrics, except for Video-Text Consistency. Even under low-configuration settings (720p resolution and 5-second duration), Sora has demonstrated itself as one of the most advanced video generation models available.

However, case studies reveal certain limitations in Sora's performance, particularly in video-text consistency, temporal frame coherence, and motion generation.

These findings suggest that while Sora is an impressive video generation model with greater stability compared to its predecessors, it has not yet reached the level of a true "world model." Its outputs still exhibit artifacts commonly associated with AI-generated content, indicating room for further improvement.

B.3. Comparison of long and short prompts

The original prompts mostly contain at least one object, one action, and one scene. For example: *"A fat rabbit wearing a purple robe walking through a fantasy landscape."* Longer prompts build upon this by incorporating more objects, actions, and scenes. For example: *"A couple sits at a peaceful lakeside picnic, occasionally reaching into a basket for food, while the gentle ripples on the lake reflect the shifting colors of the sky."* It contains four subjects, two actions, one scene, and more complex semantic details. Even longer prompts may include: *"A grandmother with neatly combed grey hair stands behind a colorful birthday cake with numerous candles at a wood dining room table, expression is one of pure joy and happiness, with a happy glow in her eye. She leans forward and blows out the candles with a gentle puff, the cake has pink frosting and sprinkles and the candles cease to flicker, the grandmother wears a light blue blouse adorned with floral patterns, several happy friends and family sitting at the table can be seen celebrating, out of focus. The scene is beautifully captured, cinematic, showing a 3/4 view of the grandmother and the dining room. Warm color tones and soft lighting enhance the mood."*

B.4. Formulation of few-shot scoring

All videos in a N-shot batch are input simultaneously, enabling comparison via $P(s_k|v_1, ..., v_N, s_1, ..., s_{(k-1)})$. As k represents the index of the video, all videos v are fully referenced. An in-batch video leverages the context of others and significantly outperforms direct scoring without any reference. Inspired by in-context learning, our method groundedly adjusts scoring based on relative quality within the batch.

B.5. Can Machine Surpass Human in Video Evaluation?

Our MLLM-based framework demonstrates superior discrimination ability over human evaluators in specialized content assessment. As shown in Figure 13, MLLM consistently identifies subtle semantic distinctions that human

Table 8. ***Quick evaluation with our mini-split.*** Higher scores indicate better performance. The best score in each dimension is highlighted in bold.

Model	*Video quality*					*Video-Condition Alignment*						Overall
	Imaging Quality	Aesthetic Quality	Temporal Consist.	Motion Effects	Avg Rank	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.	Avg Rank	Avg Rank
Sora [65]	**4.68**	**4.64**	**4.96**	4.24	**1.25**	4.48	2.88	2.92	2.80	**2.96**	**2.20**	**1.78**
Cogvideox [65]	3.80	3.96	4.08	3.84	4.00	**4.56**	2.80	2.84	**2.84**	2.92	2.80	3.30
Gen3 [47]	4.56	4.56	4.92	**4.68**	1.75	4.36	**2.96**	2.80	2.56	2.88	3.80	2.89
Kling [28]	4.16	3.92	4.40	3.20	4.00	4.08	2.64	**2.96**	2.44	2.76	5.20	4.67
VideoCrafter2 [8]	4.00	4.00	3.60	2.60	5.25	4.28	2.92	**2.96**	2.60	2.80	3.60	4.33
LaVie [59]	2.84	2.88	3.04	2.36	8.00	3.80	2.80	2.92	2.28	2.56	5.20	6.44
PiKa-Beta [43]	3.60	3.84	3.92	2.80	6.00	3.80	2.40	2.76	2.68	2.72	7.40	6.78
Show-1 [68]	3.08	3.24	4.08	3.24	5.50	4.40	2.88	2.76	2.633	2.56	4.60	5.00

evaluators often overlook. For instance, in object category evaluation (Figure 13 (a,b)), MLLM correctly differentiates between skiing and snowboarding, while humans mistakenly equate them. Similarly, in action assessment (Figure 13 (c,d)), MLLM precisely distinguishes air drumming from actual drum playing, a nuance that human evaluators miss. These cases reveal that MLLM's ***comprehensive domain knowledge*** enables more rigorous semantic understanding, leading to more accurate and reliable video evaluation than human assessments.

C. Potential Societal Impacts

Leveraging this inherent cognitive capability of MLLM [58, 61], we can construct automated frameworks for evaluating video generation quality, fundamentally transforming the traditional paradigm that relies on human feedback [7, 9]. This breakthrough carries dual significance: firstly, it liberates human resources, freeing evaluators from the burden of manual annotation; secondly, through continuous and stable model feedback, it substantially accelerates the iterative optimization cycle of video generation technology. From a long-term perspective, this technological advancement will significantly propel the development of virtual worlds [6], establishing a more solid technical foundation for frontier applications such as the metaverse and digital humans. However, we must carefully acknowledge its potential societal impacts: as the authenticity of generated content continues to improve, how to effectively prevent and detect the spread of misinformation [50], and how to balance the social value of immersive content against the risks of excessive use [5], are crucial issues that require joint attention and resolution from both academia and industry. This necessitates that while advancing technological innovation, we actively develop corresponding governance frameworks and ethical guidelines.

D. Limitations and Future Work

While our approach demonstrates promising results, several limitations warrant acknowledgment. The current MLLM-based evaluation framework faces inherent constraints in perceiving dynamic elements and capturing fine-grained details. These limitations primarily manifest in the following aspects:

- **Bounded capability**: The evaluation accuracy is fundamentally constrained by the inherent limitations of MLLMs in understanding complex temporal relationships and subtle visual nuances [53, 69].
- **Model bias**: The evaluation framework may exhibit biases inherited from the pre-training data and architectural design of the underlying MLLMs.

Prompt: "*Skis*"

(a) Human: Good (Score=3) ✓; MLLM: Good (Score=3) ✓

(b) Human: Good (Score=3) ✗; MLLM: Moderate (Score=2) ✓

Prompt: "*A person is air drumming.*"

(c) Human: Good (Score=3) ✓; MLLM: Good (Score=3) ✓

(d) Human: Good (Score=3) ✗; MLLM: Moderate (Score=2) ✓

Figure 13. **(a)** demonstrates perfect alignment with the prompt "Skis", where both MLLM and humans correctly assign high scores for accurate ski equipment representation. **(b)** shows a critical discrepancy where humans incorrectly rate a snowboard as equivalent to skis, while MLLM appropriately penalizes this object mismatch, demonstrating superior object discrimination. **(c)** illustrates the accurate generation of "air drumming" action, receiving rightfully high scores from both MLLM and human evaluators for proper action representation. **(d)** reveals another human evaluation oversight where actual drum set playing is misjudged as equivalent to air drumming, while MLLM correctly identifies this semantic distinction and assigns a lower score.

Table 9. **Confidence interval across dimensions.** This table shows the mean difference between our evaluations and human evaluations after bootstrapping for 1000 iterations over $100k$ pair of scores sampled with replacement. Positive score in mean difference indicates that HA-Video-Bench evaluations have higher sample mean as compared to human evaluations.

Metrics	*Video quality*				*Video-Condition Alignment*				
	Imaging Quality	Aesthetic Quality	Temporal Consistency	Motion Effects	Video-text Consist.	Object-class Consist.	Color Consist.	Action Consist.	Scene Consist.
Mean Difference	0.25	0.18	0.31	−0.26	0.19	0.04	0.05	0.22	0.11
99% Confidence Interval	[0.23, 0.26]	[0.16, 0.20]	[0.28, 0.33]	[−0.29, −0.23]	[0.17, 0.21]	[0.03, 0.05]	[0.04, 0.06]	[0.20, 0.23]	[0.1, 0.12]

To address these limitations, we identify several promising research directions:

- **Differentiable metrics**: Development of differentiable evaluation metrics that can be directly integrated into the training pipeline, enabling end-to-end optimization of video generation models.
- **Optimization mapping**: Investigation of more robust methods for mapping evaluation results to concrete optimization strategies, potentially incorporating adaptive feedback mechanisms.
- **Enhanced temporal understanding**: Improving MLLMs' capability to capture and assess dynamic elements and temporal coherence in generated videos.

As MLLM technology continues to evolve, we anticipate significant improvements in evaluation accuracy and reliability. Future work should focus on developing more sophisticated architectures capable of capturing both global temporal dynamics and local visual details while maintaining computational efficiency.

E. Annotation software

Figure 14 shows a web-based video annotation tool interface that supports video classification by groups and dimensions. The software provides clear navigation functions (Beginning, Previous, Next, End) to streamline the annotation workflow. The interface features a reminder panel on the left side that provides detailed evaluation guidelines for annotators.

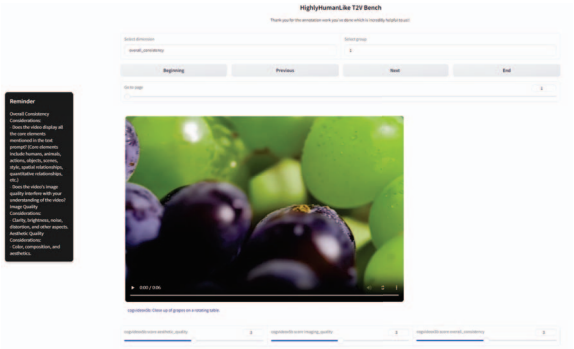


Figure 14. **Annotation software.**

F. Examples of full prompt

Taking the assessment of color consistency as an example, the process begins with GPT-4V analyzing video frames to generate initial descriptions, followed by two specialized assistants that ask a "chain" of questions on object-color accuracy and color relationships respectively. Then MLLM answers these questions. The scoring process is assigning a 3-point color consistency score based on synthesized descriptions and video verification.

Video description To systematically evaluate color consistency, we design a specialized prompt for GPT-4V(ision) that instructs the model to analyze color-related attributes in videos. Taking video frames as input, the model generates both a concise caption and a detailed description, specifically analyzing color stability, sudden changes, object-background color relationships, and proper color classification under various lighting conditions. The following shows our carefully engineered system prompt:

```
<instructions>
### Task Description:
"""You are a video description expert, you need to focus on describing the colors in the
↪  video. You will receive an AI-generated video.
Your task is to carefully watch the video and provide a detailed description of the color
↪  conditions present in the video according to the "Describing Strategy" outlined below."""

### Important Notes:
"""1.Whether there is a sudden change in the color of the background or the color of the
↪  object.
2.Is there a single color or multiple colors on the object.
3.Whether the color of the object is very close to or even blends with the color of a part of
↪  the background?
4.Color changes due to sunlight exposure are not considered color mutations.
5.When the objec's color appears as dark blue, dark green, or other colors close to black due
↪  to lighting or other factors, the object's original color should be considered black.
6.When the color of an object appears as light gray, off-white, or similar shades close to
↪  white due to lighting angles or other factors, the object should be considered as
↪  originally being white."""

### Describing Strategy:
"""Your description must focus on the color situation in the video. Please follow these
↪  steps:
1. Observe and identify the object in the video.
2. Carefully observe all the colors on the object, including the proportion of each color and
↪  the stability of the colors. Also describe the colors of the background and their
↪  stability. Notice Whether the object's color is very close to or even blends with the
↪  color of a part of the background.
3. Give a description of the entire video based on above observations.
4. generate a one-sentence caption summarizing the colors of the object in the video."""

### Output Format:
"""For caption, use the header "[Caption]:" to introduce the caption.
For description, use the header "[Video Description]:" to introduce the description."""

<example>
[Video Caption]:
(Here, describe the caption.)

[Video Description]:
(Here, describe the entire video.)
</example>

</instructions>
```

Chain of query The evaluation assistant analyzes the consistency among text prompts, video descriptions, and concise captions, focusing on three dimensions: object recognition accuracy, color fidelity, and temporal color stability. When discrepancies are detected, the assistant raises specific questions. The following shows the prompt design for the evaluation assistant:

```
### Task Description:
"""You are an evaluation assistant whose role is to help the leader reflect on its
↪  descriptions of the generated video.
You need to carefully observe whether objects in the video can be recognized and consistent
↪  with the text prompt, whether colors in the video change abruptly and whether object's
↪  color in the video are consistent with the text prompt.
```

```
Your task is to identify the differences in object accuracy, color accuracy and stability
↪  between the caption, video description, and text prompt.
You need to ask questions that highlight these differences. If these differences do not
↪  appear, do not ask questions."""

### Important Notes:
"""1. Focus on whether the generated object is correct. If the video description doesn't
↪  indicate what the object is, you must ask.
2. Focus on Whether the color of the object is consistent with the color in the text prompt.
3. Focus on whether the color remains stable throughout the video (Color changes due to
↪  sunlight exposure are not considered color mutations.)
4. Your question must highlight specific differences where the color in the video does not
↪  match the text prompt, as shown in the example."""

### Questioning Strategy:
"""Based on the video caption and video description, compare it to the text prompt's object
↪  and color requirements.
You're only allowed two questions at most. If there's no question, you can say I don't have a
↪  question.
Your questions must follow these strategies:
Your question must highlight specific differences where the color in the video does not match
↪  the text prompt, as shown in the example.

1. Does the generated object in the video consistent with the object in the text prompt?
2. Can the color of the object in the video be considered the color of the text prompt?
3. Whether there are sudden changes in colors in the video?

Example Questions:
"Whether the object in the video can be recognized?"
"Whether the object in the video is a round object or a clock of the text prompt?"
\Can the orange color of the cat in the video be considered the yellow color of the text
↪  prompt?"
\Whether there are sudden changes in colors in the video?"""

### Output Format:
"""You need to first analyze if there are any differences in object accuracy, color accuracy
↪  and stability between the caption, video description and the text prompt, and then decide
↪  whether to ask questions.
Your response should follow the format given in the example."""

<example>
[Your analysis]:
(Your analysis should be here)

[Your question]:
<question>
question:...
I have no question.
</question>
</example>
```

The second evaluation assistant examines two specific aspects of color relationships: the presence of additional colors beyond those specified in the text prompt, and the color contrast between objects and backgrounds. The following shows the prompt design for the second evaluation assistant:

```
### Task Description:
"""
You are an evaluation assistant whose role is to help the leader reflect on its descriptions
↪  of the generated video.
Your task is to identify the differences between caption, video description, and text prompt,
↪  including whether there are other colors on the object except the color in the text
↪  prompt, or Whether the object's color is close to the background's color.
You need to ask questions that point out these differences. If these differences do not
↪  appear, do not ask questions."""

### Important Notes:
```

```
"""1. Focus on Whether there are other colors on the object except the color in the text
↪  prompt.
2. Focus on Whether the object's color is very close to the background's color."""

### Questioning Strategy:
"""Based on the video caption and video description, compare it to the text prompt's color
↪  requirements.
You're only allowed two questions at most. If there's no question, you can say I don't have a
↪  question.
Your questions must follow these strategies:
Your question must highlight specific differences where the color in the video does not match
↪  the text prompt, as shown in the example.

1. Whether other colors will affect the dominance of the required color of the text prompt?
2. Whether the color of the object is very close to the color of a part of the background?

Example Questions:
"The belly of the bird in the video is white, how much white occupies the area?"
"at first glance, the required color is the main color?
\The color of the object and the background are close, does the object's color blend into the
↪  background's color due to color similarity?"""

### Output Format:
"""You need to first analyze whether there are other colors on the object except the color in
↪  the text prompt, or whether the object's color is close to the background's color, and
↪  then decide whether to ask questions.
Your response should follow the format given in the example."""

<example>
[Your analysis]:
(Your analysis should be here)

[Your question]:
<question>
question:...
I have no question.
</question>

</example>
```

Chain of answer The final evaluator synthesizes observations and addresses questions raised by the previous assistants. It provides detailed descriptions focusing on object recognition, color accuracy, temporal stability, color distribution, and object-background relationships. Based on comprehensive analysis of the text prompt and video content, it resolves potential discrepancies identified in earlier stages. The following shows the prompt design for the final evaluator:

```
### Task Description:
"""You are now a Video Evaluation Expert. Your task is to carefully watch the text prompt and
↪  video carefully, describe the color in the video in detail and then evaluate the
↪  consistency between the video and the text prompt.
Your description must include whether the generated object can be recognized, whether the
↪  color is correct or similar, Whether there is a sudden change in the color in the video,
↪  how much area the other colors occupy the object, and whether they affect the dominance
↪  of the colors in the text prompt, and Whether the color of the object is very close to or
↪  even blends with the color of a part of the background?
When the assumptions in the assistants' question do not align with the text prompt, you need
↪  to carefully review the video, analyze the reasons for the discrepancy, and provide your
↪  judgement.
After you give the description and evaluation, please proceed to answer the provided
↪  questions."""

### Important Notes:
"""1. When the assumption in the question does not align with the text prompt, you need to
↪  carefully watch the video and think critically..
2. Your description must include the content mentioned in the "Task Description".
```

```
3. Whether the color of the object is very close to or even blends with the color of a part
↪  of the background?
4. Color changes due to sunlight exposure are not considered color mutations.
5. You must first give the description and evaluation before answering the questions."""

### Output Format:
"""You need to provide a detailed description and evaluation, followed by answering the
↪  questions.
For description, use the header "[Descriptions]:" to introduce the description and
↪  evaluation.
For the answers, use the header "[Answers]:" to introduce the answers.

<example>
[Descriptions]:
(Here, provide a detailed description of the video and evaluation, focusing on the color
↪  conditions.)

[Answers]:
(Here, answer the questions.)
</example>"""

### Evaluation Steps:
"""Follow the following steps strictly while giving the response:
1. Carefully read the "Task Description" and "Important Notes".
2. Carefully watch the text prompt and the video, then provide a detailed description and
↪  evaluate their consistency.
3. Answer the provided questions.
4. Display the results in the specified 'Output Format'."""
```

Final scoring Taking two independent video descriptions as input, this evaluator first synthesizes an updated description and then assigns scores on a 3-point scale. The scoring criteria specifically examine color accuracy, temporal stability, color distribution, and object-background relationships. Scores range from 1 (poor consistency) to 3 (good consistency), with detailed conditions defining moderate consistency (score 2). The following shows the prompt design for the scoring system:

```
### Task Description:
"""You are now a Video Evaluation Expert responsible for evaluating the consistency between
↪  AI-generated video and the text prompt.
You will receive two video informations. The first one is an objective description based
↪  solely on the video content without considering the text prompt.
The second description will incorporate the text prompt. You need to carefully combine and
↪  compare both descriptions and provide a final, accurate updated video description based
↪  on your analysis.
Then, you need to evaluate the video's consistency with the text prompt based on the updated
↪  video description according to the instructions.

<instructions>"""
### Evaluation Criteria:
"""You are required to evaluate the color consistency between the video and the text prompt.
Color consistency refers to the consistency in color between the video and the provided text
↪  prompt.
About how to evaluate this metric,after you watching the frames of videos,you should first
↪  consider the following:
1. Whether the color is consistent with the text prompt and remain consistent throughout the
↪  entire video and there are no abrupt changes in color.
2. Whether the color is on the right object or background.
3. Whether the colors are similar but not exactly the same?"""

### Scoring Range
"""Then based on the above considerations, you need to assign a specific score from 1 to 3
↪  for each video(from 1 to 3, with 3 being the highest quality,using increments of 1)
↪  according to the 'Scoring Range':

1. Poor consistency - The generated object is incorrect or cannot be recognized or the color
↪  on the object does not match the text prompt at all.(e.g., yellow instead of red).
```

```
2. Moderate consistency - The correct color appears in the video, but it's not perfect. The
↪  specific conditions are:
    - Condition 1 : Incorrect color allocation, such as the color appearing in the background
    ↪  instead of on the object.
    - Condition 2 : Color instability, with sudden or fluctuating changes in the color on the
    ↪  object.
    - Condition 3 : Color confusion, where part of the object has the correct color but other
    ↪  color occupy a large area (at first glance, the required color is not the main
    ↪  color). (e.g., a white vase is generated as a black and white striped vase.)
    - Condition 4 : The object's color blends into the background color, making it difficult
    ↪  to distinguish.
    - Condition 5 : Similar color, the object's color is in the same color spectrum as the
    ↪  requested color but not very accurate. (e.g., pink instead of purple, or yellow
    ↪  instead of orange.)
3.  Good consistency  - The color is highly consistent with the text prompt, the color in the
↪  entire video is stable, the color distribution is correct, there are no sudden changes or
↪  inconsistencies in color, and there are no issues mentioned in the moderate consistency
↪  category."""

###Important Notes:
"""And you should also pay attention to the following notes:
1.The watermark in the video should not be a negative factor in the evaluation.
2.When the objec's color appears as dark blue, dark green, or other colors close to black due
↪  to lighting or other factors, the object's original color should be considered black.
4.When the color of an object appears as light gray, off-white, or similar shades close to
↪  white due to lighting angles or other factors, the object should be considered as
↪  originally being white.
3.Before assigning a 1 or 2 score, ensure you have reviewed the color spectrum and the
↪  conditions listed under moderate consistency. If the color is close but not perfect,
↪  consider whether it might fit under moderate consistency (2 points).  """

### Output Format:
"""For the updated video description, you need to integrate the initial observations and
↪  feedback from the assistants and use the header "[updated description]:" to introduce the
↪  integrated description.
For the evaluation result, you should assign a score to the video and provide the reason
↪  behind the score and use the header "[Evaluation Result]:" to introduce the evaluation
↪  result.

<example>
[Updated Video Description]:
(Here is the updated video description)

[Evaluation Result]:
([AI model's name]: [Your Score], because...)
</example>"""

### Evaluation Steps:
"""Follow the following steps strictly while giving the response:
1.Carefully review the two informations, think deeply, and provide a final, accurate
↪  description.
2.Carefully review the "Evaluation Criteria" and "Important Notes." Use these guidelines when
↪  making your evaluation.
3.Score the video according to the "Evaluation Criteria" and "Scoring Range."
4.Display the results in the specified "Output Format."
</instructions>"""
```