OSCBench: Benchmarking Object State Change in Text-to-Video Generation

Anonymous ACL submission

Abstract

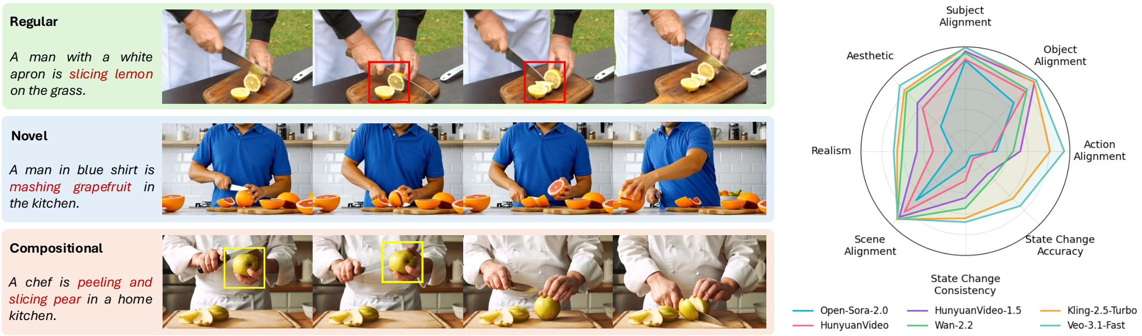
Text-to-video (T2V) generation models have made rapid progress in producing visually high-quality and temporally coherent videos. However, existing benchmarks primarily focus on perceptual quality, text–video alignment, or physical plausibility, leaving a critical aspect of action understanding largely unexplored: object state change (OSC) explicitly specified in the text prompt. OSC refers to the transformation of an object's state induced by an action, such as peeling a potato or slicing a lemon. In this paper, we introduce OSCBench, a benchmark specifically designed to assess OSC performance in T2V models. OSCBench is constructed from instructional cooking data and systematically organizes action–object interactions into regular, novel, and compositional scenarios to probe both in-distribution performance and generalization. We evaluate six representative open-source and proprietary T2V models using both human user study and multimodal large language model (MLLM)–based automatic evaluation. Our results show that, despite strong performance on semantic and scene alignment, current T2V models consistently struggle with accurate and temporally consistent object state changes, especially in novel and compositional settings. These findings position OSC as a key bottleneck in text-to-video generation and establish OSCBench as a diagnostic benchmark for advancing state-aware video generation models.

1 Introduction

Text-to-video (T2V) generation models have made remarkable progress in recent years, producing videos with increasingly high visual fidelity and temporal coherence. These advances have enabled a wide range of applications, including creative content generation, instructional video synthesis, and simulation of real-world processes (Google DeepMind, 2025b; Ma et al., 2025). As T2V models continue to scale, a central question emerges: to what extent do these models faithfully realize the consequences of actions specified in language, rather than merely producing visually appealing motion patterns?

Recent benchmarks have taken important steps toward answering this question by evaluating physical plausibility and commonsense constraints in generated videos, such as adherence to gravity, collisions, and material properties (Meng et al., 2024; Gu et al., 2025). While these evaluations probe fundamental aspects of physical realism, they overlook a critical dimension of language-grounded action understanding that is ubiquitous in everyday activities: Object State Change (OSC) explicitly specified by the prompt. In many real-world tasks, such as slicing a lemon, peeling a carrot, or mixing dough, success is defined not only by performing an action, but by transforming an object from an initial state to a specific target state (e.g., a whole lemon becoming sliced). Correctly modeling such object state change is essential for downstream applications, including robotics, embodied AI, and instructional video generation.

Object state change poses a particularly stringent test of language-grounded reasoning in T2V models. Correct OSC generation requires a model to understand the action semantics expressed in language, infer the intended object transformation, and render a continuous and coherent visual evolution over time. However, despite producing visually compelling videos, current T2V models often fail on this dimension: generated outputs may appear realistic at a glance while exhibiting incorrect, incomplete, or inconsistent object state changes. Figure 1 (a) illustrates representative failure cases, where objects change into implausible states or the instructed action is misunderstood, revealing a gap between high-level semantic alignment and faithful realization of action consequences. Despite the importance of OSC, it has not been systematically evaluated in existing T2V benchmarks,



(a) (b)

Figure 1: Overview of OSCBench evaluation. (a) Representative failure cases from regular, novel, and compositional object state change scenarios. In the regular case, the red box marks an implausible state change of the lemon during slicing. In the novel case, the model misinterprets the instructed action, resulting in a wrong object transformation. In the compositional case, the yellow box indicates an incomplete state change where the pear remains unpeeled. (b) Human-evaluated multi-dimensional performance of T2V models on OSCBench.

which primarily emphasize overall perceptual quality, text–video alignment, or physical plausibility, without explicitly assessing whether an object reaches the the correct target state or whether the state transition unfolds consistently over time.

To address this gap, we introduce OSCBench, a benchmark designed to evaluate object state change in text-to-video generation. We focus on instructional cooking scenarios, where state changes are frequent, diverse, and well-defined, and build OSCBench on top of the HowToChange dataset (Xue et al., 2024). To enable balanced and comprehensive evaluation, we abstract actions and objects into semantically meaningful categories and then construct three complementary evaluation regimes as shown in Figure 1 (a): regular scenarios covering common action–object pairs (e.g., slicing lemon), novel scenarios that test generalization to uncommon yet feasible state changes (e.g., mashing grapefruit), and compositional scenarios involving multiple action compositions (e.g., peeling and slicing pear). In total, OSCBench comprises 1,120 prompts across 140 object-state scenarios, providing a specific benchmark for evaluating OSC performance in T2V models.

In addition, we evaluate six state-of-the-art (SOTA) T2V models on OSCBench, including four widely used open-source systems (Open-Sora-2.0 (Peng et al., 2025), HunyuanVideo (Kong et al., 2024), HunyuanVideo-1.5 (Team, 2025), Wan-2.2 (Wan et al., 2025)) and two proprietary models (Kling-2.5-Turbo (KlingAI, 2025) and Veo-3.1-Fast (Google DeepMind, 2025b)). We conduct both human user study and automatic evaluation using the latest multimodal large language models (MLLMs). Across the two evaluation methods, we design a comprehensive set of criteria covering semantic adherence, OSC performance, scene alignment, and perceptual quality. In particular, rather than using MLLMs as black-box scorers, we employ Chain-of-Thought (Wei et al., 2022) evaluation strategy that explicitly guides the reasoning process through criteria grounding, evidence extraction, and score justification. We further analyze the correlation between human judgments and MLLM-based evaluations to assess the reliability of automated OSC evaluation. Our results in Figure 1 (b) reveal that while SOTA T2V models generally perform well on high-level semantic alignment (e.g., subject, object, and scene), object state change accuracy and consistency remain a significant challenge. These findings position OSC as a critical diagnostic dimension that complements existing evaluations. By revealing how state changes deviate from intended action effects, OSCBench provides practical guidance for building video generation models that reason more faithfully about actions and their consequences.

In summary, our contributions are three-fold:

- We introduce OSCBench, the first benchmark explicitly designed to evaluate object state change in text-to-video generation across regular, novel and complex scenarios.
- We design a set of criteria covering semantic adherence, OSC performance, scene alignment, and perceptual quality to comprehensively evaluate the video generation performance with both human user study and automatic MLLM assessment.
- We benchmark six SOTA T2V models, sys-

tematically examine their performance across different OSC scenarios, and identify key challenges that persist. The results offer guidance for designing models with OSC-aware generation and outline directions for future research.

2 Related Work

Benchmarks for Text-to-Video Generation. The rapid advancement of T2V models has motivated the development of benchmarks for accurate and reliable assessment. A number of recent benchmarks (Huang et al., 2024; He et al., 2024) aim to provide systematic evaluation of T2V models either from a comprehensive perspective or through specific aspects of generation quality. For example, VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024) target holistic evaluation across multiple interpretable dimensions, including temporal consistency, motion smoothness, and text–video alignment. To better diagnose particular modeling challenges, several aspect-specific benchmarks have been proposed. For example, T2V-CompBench (Sun et al., 2025) evaluates compositional generation capabilities, while DEVIL (Liao et al., 2024) focuses on the dynamic characteristics of generated videos. More recently, researchers have observed that T2V models frequently generate videos that violate physical constraints. This has motivated the development of benchmarks that explicitly assess physical plausibility, such as VideoPhy (Bansal et al., 2024), PhyGenBench (Meng et al., 2024), and PhyWorldBench (Gu et al., 2025), which examine whether generated videos adhere to basic physical commonsense.

Despite these advances, existing benchmarks pay limited attention to OSC. In this work, we introduce a benchmark specifically for object state change, providing scenarios that require accurate state modeling and enabling targeted evaluation of a model's OSC understanding.

Evaluation Methods for Text-to-Video Models. Recent video benchmarks (Huang et al., 2024; Meng et al., 2024; Gu et al., 2025) commonly adopt a hybrid evaluation protocol that combines automatic model evaluation with human user study. For automatic evaluation, CLIP (Xue et al., 2024) and ViCLIP (Wang et al., 2023) based text–video similarity models are widely used to assess semantic alignment between prompts and generated videos. More recently, MLLMs have demonstrated strong abilities in understanding complex visual content (Ouyang et al., 2025; Zhang et al., 2025; He et al., 2025). Therefore, many video benchmarks (Feng et al., 2025; Motamed et al., 2025; Han et al., 2025) employ MLLMs to evaluate the semantic consistency in generated videos. Building on this capability, PhyWorldBench (Gu et al., 2025) further leverages MLLMs to evaluate whether generated videos obey physical laws, which often requires multi-step reasoning.

To evaluate fine-grained OSC, we leverage the reasoning capabilities of MLLMs and adopt a CoT strategy (Wei et al., 2022). Unlike existing benchmarks (Gu et al., 2025), which mainly use CoT to generate textual descriptions, we use it to guide models through a structured reasoning process, encouraging careful visual inspection and more reliable state-change judgments.

3 OSCBench Construction

The goal of OSCBench is to provide a structured and comprehensive benchmark for evaluating object state change in text-to-video generation. Designing such a benchmark requires addressing three key challenges: (i) covering realistic and diverse object state changes grounded in textual prompts, (ii) ensuring controlled and balanced coverage of actions and objects to reduce dataset bias, and (iii) introducing varying levels of difficulty to probe both memorization and generalization. In this section, we describe how OSCBench is constructed to meet these requirements.

3.1 Data Source and Abstraction

Object state change is ubiquitous in everyday activities, with cooking being a representative domains. Cooking tasks naturally involve diverse state transformations, such as chopping, peeling, and heating, and exhibit clear causal relationships between actions and resulting object states. We therefore build OSCBench on the HowToChange dataset (Xue et al., 2024), which is derived from instructional cooking videos in HowTo100M (Miech et al., 2019). HowToChange contains 20 fine-grained action elements and 134 object elements, yielding 409 distinct action–object combinations (e.g., slicing apple). However, these combinations exhibit a strong long-tail distribution: common pairs appear frequently (e.g., chopping potato), while many plausible ones are rare or absent (e.g., squeezing ginger). Directly sampling from this distribution would bias evaluation toward frequent

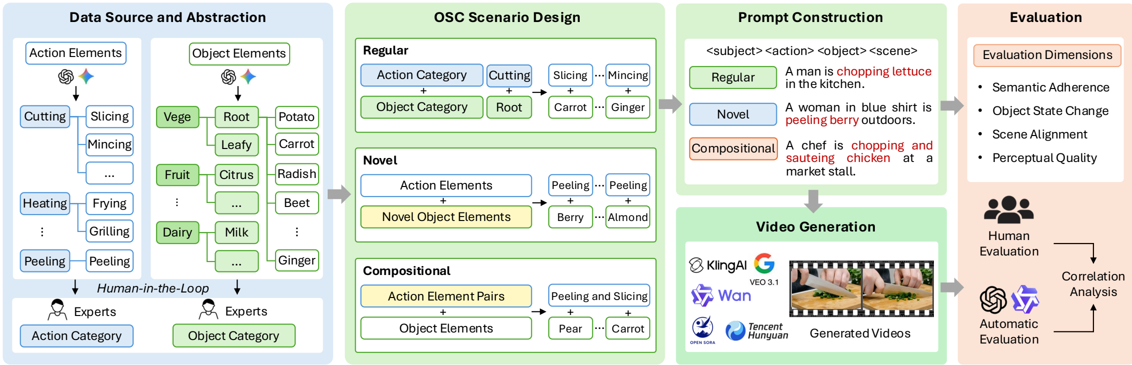


Figure 2: Overview of the OSCBench construction and evaluation pipeline. We build unified action and object categories from instructional cooking data via a human-in-the-loop process, and construct regular, novel, and compositional OSC scenarios as text prompts for video generation. The generated videos are evaluated by humans and MLLMs across multiple criteria, and we analyze their correlations to assess automatic evaluation reliability.

patterns and limit insights into generalization.

To mitigate this issue, we reorganize the raw action and object elements into high-level conceptual categories using a human-in-the-loop abstraction process. Specifically, as shown in Figure 2, guided by cooking objectives, we first use GPT-5.2 (OpenAI, 2025) and Gemini-3 (Google DeepMind, 2025a) to propose candidate groupings of the 20 action elements into 9 action categories (e.g., heating), and to cluster the 134 object elements into 8 major object categories (e.g., vegetable) with 28 finer-grained subcategories (e.g., root vegetables). These groupings are then iteratively refined and validated by human experts to ensure semantic correctness and practical plausibility. This abstraction enables systematic scenario construction while preserving semantic diversity.

3.2 OSC Scenario Design

Based on the abstracted action and object taxonomy, we construct three complementary types of OSC scenarios, designed to evaluate different aspects of OSC-aware video generation: regularity, generalization, and compositionality.

Regular OSC Scenarios. Regular scenarios are designed to cover a broad range of realistic and commonly occurring object state changes. We pair each action category with compatible object subcategories to form candidate scenarios. All candidates are first filtered using automated checks by ChatGPT and then validated by human review. This process yields 108 regular OSC scenarios. For each scenario, we further enumerate concrete instances by pairing specific action elements with object elements and manually select eight representative action–object combinations (e.g., mincing ginger), ensuring diversity while maintaining feasibility.

Novel OSC Scenarios. To evaluate whether models can reason about unfamiliar yet plausible object state changes, we introduce novel scenarios that deliberately deviate from common action-object combinations. For each of the 20 action elements, we select 8 uncommon yet feasible objects (e.g., peeling berries), resulting in 20 novel scenarios. These scenarios cannot be reliably solved through memorization of frequent action–object pairs and instead require models to infer state changes from action semantics.

Compositional OSC Scenarios. Real-world activities often involve multiple actions applied sequentially, where state changes evolves over time. To assess whether models can maintain coherent intermediate and final states, we construct compositional scenarios by composing pairs of action elements (e.g., peeling followed by slicing). We select 12 common action pairs, verified by human inspection, and combine each pair with eight suitable objects (e.g., peeling and slicing potato). These scenarios explicitly examine multiple action composition and temporal consistency for OSC-aware video generation.

3.3 Prompt Construction

For every action–object combination in each scenario, we generate prompts using a structured template: <subject><action><object><scene>. We randomly generate three candidate prompts for each combination using GPT-5.2 and manually select the most natural one. Examples can be seen in Figure 2 (e.g., A man is chopping lettuce in the kitchen). In addition to full prompts with subjects and scenes, we further test how models respond when only object state change cues are provided. Specifically, we randomly simplify 1-2 prompts per scenario to the minimal form, <action><object>. This variant reduces contextual cues and places greater emphasis on the model's ability to infer and realize OSC directly from the action description.

3.4 Benchmark Statistics

OSCBench comprises 140 object state change scenarios in total, including 108 regular scenarios, 20 novel scenarios, and 12 compositional scenarios. Each scenario contains 8 action–object combinations, resulting in 1,120 prompts overall. The prompts are concise and descriptive, with an average length of 9.2 words, providing sufficient context while avoiding unnecessary linguistic complexity. We additionally provide a word cloud for OSCBench to illustrate the word distribution in the Appendix A. By combining structured abstraction, controlled scenario design, and multiple difficulty regimes, OSCBench enables systematic analysis of object state change performance in text-to-video models, covering both common patterns and challenging generalization cases.

4 Evaluation

Evaluating text-to-video generation models is inherently challenging, particularly when the goal is to assess object state change specified by the prompt. A reliable evaluation must verify not only whether a generated video aligns with the prompt at a semantic level, but also whether the prompt-implied object state transition is realized accurately and consistently over time. While human evaluators can naturally perform such judgments, large-scale human use study is costly and difficult to scale. Following PhyWorldBench (Gu et al., 2025), we conduct both human user study and automatic assessment using multiple large language models.

4.1 Evaluation Dimensions

We comprehensively evaluate generated videos along four complementary evaluation dimensions: semantic adherence, object state change, scene alignment, and perceptual quality.

Semantic Adherence. This dimension measures whether the core semantic entities described in the prompt are faithfully grounded in the generated video. Specifically, we evaluate three key components independently: Subject alignment to measure whether the acting subject (e.g., a man or a woman) is present and correct, object alignment to evaluate whether the manipulated object matches the prompt and action alignment to assess whether the performed action corresponds to the intended action described in the prompt.

Object State Change. This is the central dimension of OSCBench. Evaluating object state change requires reasoning about both the outcome and the temporal evolution of the object. We therefore decompose OSC evaluation into two sub-dimensions: state-change accuracy, which measures whether the object reaches the correct target state implied by the prompt (e.g., a whole apple becoming sliced), and state-change consistency, which assesses whether the transformation unfolds smoothly and coherently over time, without abrupt jumps or unnatural object appearances or unexplained appearance or disappearance of object parts.

Scene Alignment. This dimension evaluates whether the global environment in the video matches the scene description in the prompt (e.g., kitchen or market). It focuses on the background context, such as whether the video clearly occurs in a kitchen or an outdoor market, and whether the scene remains stable and coherent over time.

Perceptual Quality. This dimension measures the overall visual impression of the video and consists of two aspects: realism, which measures whether the video resembles real-world footage in terms of motion, lighting, and texture, and aesthetic quality, which reflects how visually appealing the video appears in terms of composition, color, and overall presentation.

4.2 Human Evaluation

We first conduct human user study as a strong reference to evaluate our OSCBench. As exhaustive human evaluation over all generated videos would be prohibitively costly and time-consuming, we adopt a representative sampling strategy. Specifically, to cover the full diversity of OSCBench, we sample one prompt from each of the 140 OSC scenarios, ensuring that all regular, novel, and complex scenarios are represented. For each selected prompt, we generate one video for each T2V model, resulting in 140 videos per model for human evaluation. Each video is independently rated by three human evaluator across the evaluation dimensions described in Section 4.1. To encourage fine-grained and consistent judgments, we provide a 1-5 Likert scale for each dimension. For each text–video pair, we average the three evaluator' scores to obtain the mean opinion score for each evaluation dimension. These human scores serve both as primary benchmark results and as a reference signal for validating automatic evaluation using MLLMs.

4.3 MLLM-Based Automatic Evaluation

Automatic evaluation based on text–video similarity models, such as CLIP and ViCLIP, is effective for measuring coarse semantic alignment but insufficient for assessing fine-grained object state changes and perceptual quality. MLLMs have recently shown strong capabilities in visual understanding and multi-step reasoning, which can serve as reasoning-based evaluators for video generation (Gu et al., 2025). Rather than treating MLLMs as black-box scorers, we design a CoT evaluation strategy that explicitly structures the reasoning process. For each video and each evaluation dimension, the MLLM follows a three-step procedure: (1) **Criteria grounding.** The model restates the scoring criterion of each evaluation dimension in its own words, ensuring it internalizes the scoring definition before examining the video. (2) **Evidence extraction.** The model then identifies frame-level visual evidence that is relevant to the criterion and briefly explains why these observations support its assessment. (3) **Score decision.** Based on the extracted evidence, the model assigns a discrete score from 1 to 5 and explicitly links the score to the observed evidence. We provide the detailed prompt used for MLLM evaluation in the Appendix C.

We apply this procedure to all adopted MLLMs across all evaluation dimensions. By explicitly constraining the reasoning route, the CoT evaluation strategy encourages the model to focus on fine-grained object states and their temporal evolution, rather than being distracted by salient but irrelevant visual details.

5 Evaluation Results and Analysis

5.1 Experimental Setup

We evaluate six representative state-of-the-art (SOTA) Text-to-Video generation models, including four widely used open-source systems (Open-Sora-2.0 (Peng et al., 2025), HunyuanVideo (Kong et al., 2024), HunyuanVideo-1.5 (Team, 2025), and Wan-2.2 (Wan et al., 2025)) and two proprietary models (Kling-2.5-Turbo (KlingAI, 2025) and Veo-3.1-Fast (Google DeepMind, 2025b)). Detailed video generation settings are provided in the Appendix B. For automatic evaluation, we assess the

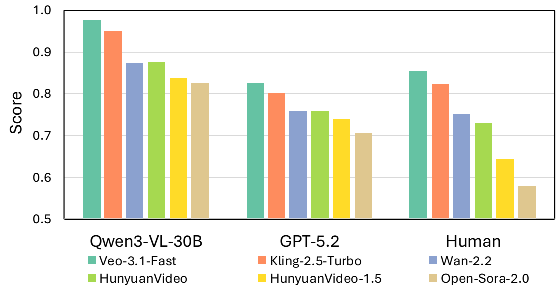


Figure 3: Overall performance comparison of T2V models based on aggregated evaluation scores from human evaluator and MLLM-based evaluators (Qwen3-VL-30B and GPT-5.2).

generated videos using ViCLIP for semantic similarity measurement as well as MLLM, including Qwen3-VL-30B, GPT-5-mini, and GPT-5.2. For space considerations, we present GPT-5.2–based evaluation results in the main paper and include results from other MLLMs in the Appendix C. All human and automatic evaluation scores are normalized to 0–1 for comparison.

5.2 Performance Comparison

Figure 3 presents the overall evaluation results by averaging the scores of all evaluation dimensions based on Qwen-VL-30B, GPT-5.2 and human judgment. Among the evaluated models, Veo-3.1-Fast achieves the strongest overall performance, followed by Kling-2.5-Turbo, while open-source models exhibit comparatively lower performance on average. Tables 1 and 2 report the human evaluation and GPT5.2-based automatic assessment scores for each model across individual evaluation dimension, respectively. Despite differences in absolute scores, both evaluation methods exhibit highly consistent trends across models. It can be observed that most models perform well on semantic adherence (particularly for subject and object) and scene alignment, but exhibit substantially lower scores on OSC accuracy and consistency. This discrepancy suggests that current T2V models are generally capable of grounding high-level semantics from text, yet struggle to faithfully model the consequences of actions on object states over time. Notably, realism also remains challenging, particularly in terms of human evaluation, suggesting that limitations in accurately modeling object state changes are often accompanied by residual visual artifacts, even when aesthetic quality is relatively strong.

To further illustrate these findings, Figure 4 shows example videos generated by different mod-

Model	Semantic Adherence			Object State Change		Scene Alignment	Perceptual Quality	
	Subject	Object	Action	Accuracy	Consistancy		Realism	Aesthetics
Open-source models								
Open-Sora-2.0 (Peng et al., 2025)	0.860	0.734	0.518	0.380	0.428	0.740	0.416	0.540
HunyuanVideo (Kong et al., 2024)	0.868	0.826	0.494	0.402	0.510	0.834	0.526	0.688
HunyuanVideo-1.5 (Team, 2025)	**0.914**	**0.902**	**0.656**	0.524	0.608	0.876	0.618	0.730
Wan-2.2 (Wan et al., 2025)	0.904	0.842	0.616	**0.560**	**0.668**	**0.894**	**0.702**	**0.818**
Proprietary models								
Kling-2.5-Turbo (KlingAI, 2025)	**0.938**	0.900	0.826	0.726	0.726	**0.894**	0.732	0.836
Veo-3.1-Fast (Google DeepMind, 2025b)	0.936	**0.916**	**0.908**	**0.786**	**0.748**	0.890	**0.752**	**0.874**

Table 1: Human evaluation results of different T2V models across multiple evaluation dimensions.

Model	Semantic Adherence			Object State Change		Scene Alignment	Perceptual Quality	
	Subject	Object	Action	Accuracy	Consistancy		Realism	Aesthetics
Open-source models								
Open-Sora-2.0 (Peng et al., 2025)	0.910	0.722	0.616	0.512	0.658	0.892	0.634	0.712
HunyuanVideo (Kong et al., 2024)	0.898	0.764	0.562	0.466	0.730	0.948	0.752	0.782
HunyuanVideo-1.5 (Team, 2025)	**0.982**	**0.788**	**0.642**	**0.546**	0.708	0.936	0.736	0.778
Wan-2.2 (Wan et al., 2025)	0.950	0.774	0.570	0.518	**0.710**	**0.974**	**0.768**	**0.798**
Proprietary models								
Kling-2.5-Turbo (KlingAI, 2025)	**0.990**	0.792	0.742	0.652	0.692	0.972	0.772	**0.802**
Veo-3.1-Fast (Google DeepMind, 2025b)	0.976	**0.834**	**0.802**	**0.740**	**0.702**	**0.978**	**0.782**	**0.802**

Table 2: GPT-5.2–based evaluation results of T2V models on OSCBench across multiple evaluation dimensions.

els for the same object-state-change prompt. In the first three models, the object state change is incorrect, where the apple is not sliced into pieces. Although videos generated by Wan-2.2, Kling-2.5-Turbo, and Veo-3.1-Fast successfully exhibit slicing behavior, they still suffer from issues in state change consistency or noticeable artifacts. For instance, Wan-2.2 shows a half-sliced apple reverting to a whole state (red box), Kling-2.5-Turbo produces unreal interactions between the knife and the bowl (yellow box), and Veo-3.1-Fast introduces an additional apple chunk in the final frame (green box). Despite these issues, most models correctly render the subject, object, and scene, reinforcing the conclusion that high-level semantic alignment is substantially easier than accurate and consistent object state change modeling.

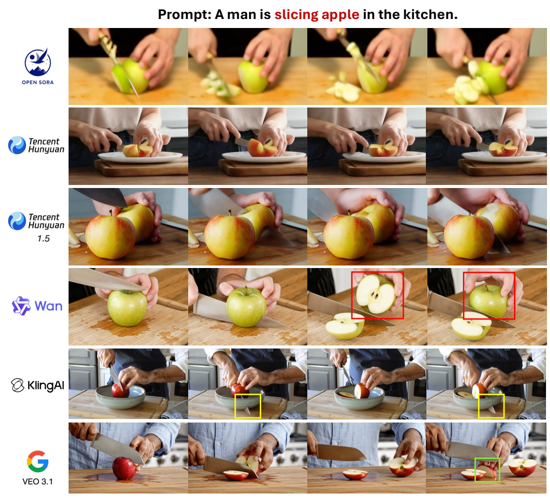


Figure 4: Sampled video frames generated by different T2V models. State change consistency or noticeable artifacts are highlighted in boxes.

5.3 Human–MLLM Correlation Analysis

We analyze the correlation between human and automatic evaluation results to assess the reliability of MLLM-based evaluation. We report the model correlations with human evaluation in terms of Kendall's τ and Spearman's ρ in Table 3, and include inter-evaluator agreement among human evaluators as a reference. Overall, MLLM-based evaluators exhibit substantially higher correlation with human judgments than the text–video similarity model ViCLIP across all evaluation dimensions, highlighting the advantage of multimodal reasoning over similarity-based scoring. Among all evaluated MLLMs, GPT-5.2 by incorporating the Chain-of-Thought (CoT) evaluation strategy generally achieves the strongest overall agreement with human evaluation, indicating that explicitly structured reasoning helps the model better identify fine-grained visual cues and state transitions. Despite these strengths, we also observe noticeably weaker correlations on perceptual-quality metrics (i.e., realism and aesthetics) for GPT-5.2 with CoT compared to other MLLMs. This gap likely reflects the inherently subjective nature of such judgments and indicates that fully automating perceptual assessment remains challenging. Besides, although human-MLLM correlations are still lower than human–human agreement, Figure 3 shows that the MLLM-based evaluation produces the same overall ranking of T2V systems as human. This consistency suggests that automatic evaluation based on MLLM, while imperfect at the fine-grained scoring,

Metrics	Semantic Adherence						Object State Change				Scene Alignment		Perceptual Quality			
	Subject		Object		Action		Accuracy		Consistancy				Realism		Aesthetics	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
ViCLIP	0.106	0.132	0.195	0.245	0.288	0.386	-	-	-	-	-	-	-	-	-	
Qwen3-VL-30B	0.406	0.413	0.412	0.429	0.542	0.624	0.426	0.503	0.289	0.341	0.145	0.149	0.269	0.297	0.407	0.426
GPT-5-mini	**0.433**	**0.439**	0.428	0.441	0.478	0.543	0.342	0.392	0.243	0.259	0.200	0.206	0.303	0.338	**0.514**	**0.541**
GPT-5.2 (w/o CoT)	0.295	0.318	0.409	0.444	0.623	0.703	0.415	0.493	0.303	0.343	0.425	0.447	**0.323**	**0.355**	0.393	0.411
GPT-5.2	0.369	0.374	**0.433**	**0.466**	**0.628**	**0.710**	**0.427**	**0.507**	**0.317**	**0.359**	**0.485**	**0.505**	0.276	0.318	0.367	0.385
Human	0.468	0.472	0.484	0.506	0.636	0.735	0.603	0.691	0.501	0.598	0.492	0.517	0.613	0.711	0.581	0.647

Table 3: Correlation between human and MLLM-based automatic evaluations in terms of Kendall's τ and Spearman's ρ. The last row reports the mean inter-human correlation for reference.

Models	Object State Change Scenario		
	Regular	Novel	Compositional
Open-source models			
Open-Sora-2.0 (Peng et al., 2025)	0.410	0.389	**0.416**
HunyuanVideo (Kong et al., 2024)	**0.472**	0.405	0.437
HunyuanVideo-1.5 (Team, 2025)	**0.572**	0.559	0.556
Wan-2.2 (Wan et al., 2025)	**0.635**	0.531	0.594
Proprietary models			
Kling-2.5-Turbo (KlingAI, 2025)	**0.744**	0.714	0.699
Veo-3.1-Fast (Google DeepMind, 2025b)	0.797	0.731	**0.805**

Table 4: Human-evaluated object state change scores of T2V models across regular, novel, and compositional scenarios, averaged over accuracy and consistency.

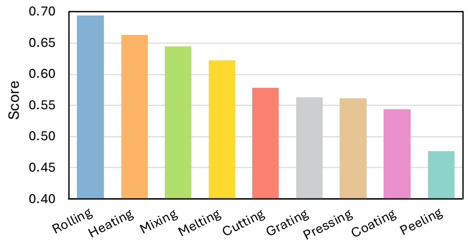


Figure 5: Object state change performance across action categories by human evaluation. Scores are averaged over accuracy and consistency.

is nevertheless reliable for assessing overall model performance trends at scale.

5.4 Category Analysis

Table 4 presents the object state change performance of different T2V models across regular, novel, and compositional scenarios. Regular scenarios mainly achieve the highest scores across all models, as they largely reflect common action–object combinations that are well represented in training data. Novel scenarios exhibit the most severe performance degradation, indicating that current T2V models struggle to generalize state-change reasoning to uncommon but feasible action–object pairs. In contrast, compositional scenarios generally perform better than novel ones but worse than regular ones. This suggests that composing multiple familiar actions in sequence is less challenging than reasoning about unseen combinations, yet still requires maintaining coherent intermediate states over time.

To further investigate how different actions affect object state change performance, Figure 5 reports results across action categories. Models achieve higher scores on relatively simple actions with clear and visually salient transformations, such as rolling and heating (e.g, rolling dough or heating root vegetables), where state changes are localized and temporally straightforward. In contrast, performance drops substantially for actions involving complex hand–object interactions or subtle visual transitions, such as peeling, coating, and pressing (e.g., peeling carrot or coating shrimp). These actions require precise manipulation and gradual appearance changes, making both the state change and its visual evidence harder to model.

6 Conclusion

We have presented OSCBench, a benchmark for evaluating text-to-video generation with a focus on object state change. OSCBench systematically characterizes regular, novel, and complex state-transition scenarios, covering a broad spectrum of cooking activities. Using this benchmark, we evaluate six representative T2V models using both human user study and MLLM-based automatic assessment, and analyze the correlation between the two methods to assess the reliability of automatic evaluation. Our experiments demonstrate that existing models generally succeed at grounding high-level semantics and producing visually appealing content, but they struggle to accurately and consistently model object state change over time. These limitations persist across regular, novel, and compositional scenarios, and are particularly pronounced for actions involving subtle or complex hand–object interactions. Overall, OSCBench, together with our evaluation framework and empirical analyses, reveals fundamental limitations of existing T2V systems in modeling object state change, and provides a diagnostic foundation for developing more state-aware and robust video generation models in future work.

Limitations

While OSCBench provides a focused benchmark for evaluating object state change in text-to-video generation, it has several limitations. First, OSCBench primarily focuses on cooking-related manipulation scenarios, which offer clear and well-defined object state changes but do not fully capture the diversity of interactions found in other domains, such as tool use, household assembly, or outdoor activities. Although cooking covers a wide range of everyday manipulations, extending OSCBench to broader domains would further improve its generality and applicability. Second, our evaluation emphasizes comparative and diagnostic analysis rather than exhaustive human annotation of all generated videos, due to practical cost and scalability constraints. While our sampling strategy ensures balanced coverage across regular, novel, and compositional scenarios, larger-scale human evaluation could reveal additional fine-grained failure modes that are not fully captured in the current setting. We view these limitations as opportunities for future work and hope that OSCBench will serve as a foundation for extending object state change evaluation to broader domains and more comprehensive assessment protocols.

Ethical Considerations

Our study employs human evaluation to assess video generation quality and to serve as a reference for validating the reliability of MLLM-based automatic scoring. A representative subset of generated videos was rated by human according to clearly defined criteria. Participants were informed about the study and provided informed consent prior to participation. Since the task involved only the evaluation of model-generated videos, no personal or sensitive information was collected.

The evaluation tasks did not expose participants to harmful or sensitive content. All prompts used in the benchmark were reviewed by the authors to ensure that no unsafe or dangerous material was included. Our work is conducted solely for research purposes and aims to improve the reliability and transparency of multimodal evaluation, rather than to create or promote harmful applications.

References

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. 2024. Videophy: Evaluating physical commonsense for video generation. In *The Thirteenth International Conference on Learning Representations*.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. 2025. Tc-bench: Benchmarking temporal compositionality in conditional video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4638–4662.

Google DeepMind. 2025a. Gemini 3.

Google DeepMind. 2025b. Veo 3.1.

Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and 1 others. 2025. Phyworldbench: A comprehensive evaluation of physical realism in text-to-video models. *arXiv preprint arXiv:2507.13428*.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, and 1 others. 2025. Video-bench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18858–18868.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, and 1 others. 2024. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123.

Zhitao He, Sandeep Polisetty, Zhiyuan Fan, Yuchen Huang, Shujin Wu, and Yi R Fung. 2025. Mmboundary: Advancing mllm knowledge boundary awareness through reasoning step confidence calibration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16427–16444.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818.

KlingAI. 2025. Klingai.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, and 1 others. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Mingxiang Liao, Qixiang Ye, Wangmeng Zuo, Fang Wan, Tianyu Wang, Yuzhong Zhao, Jingdong Wang, Xinyu Zhang, and 1 others. 2024. Evaluation of text-to-video generation models: A dynamics perspective. *Advances in Neural Information Processing Systems*, 37:109790–109816.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. 2024. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, and 1 others. 2025. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. 2024. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. 2025. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*.

OpenAI. 2025. Gpt-5.2.

Kun Ouyang, Yuanxin Liu, Shicheng Li, Yi Liu, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2025. Punchbench: Benchmarking mllms in multimodal punchline comprehension. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 986–1008.

Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, and 13 others. 2025. Open-sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. 2025. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8406–8416.

Tencent Hunyuan Foundation Model Team. 2025. Hunyuanvideo 1.5 technical report. *Preprint*, arXiv:2511.18870.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, and 43 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, and 1 others. 2023. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zihui Xue, Kumar Ashutosh, and Kristen Grauman. 2024. Learning object state changes in videos: An open-world perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18493–18503.

Zicheng Zhang, Xiangyu Zhao, Xinyu Fang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, Haodong Duan, Kai Chen, and Guangtao Zhai. 2025. Redundancy principles for MLLMs benchmarks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12504. Association for Computational Linguistics.

In this appendix, we first present a detailed overview of OSCBench in Section A and describe the video generation settings in Section B. Section C provides the evaluation procedure and additional MLLM-based results. Finally, Section D offers illustrative examples of different object state change scenarios, together with analyses that facilitate a deeper understanding of the benchmark.

A OSCBench Details

Data Abstraction Results. In OSCBench contruction, we begin with categorizing the actions and objects in the HowToChange dataset. The taxonomy is constructed through GPT-5.2–assisted grouping, cross-checked with Gemini-3, and subsequently subjected to human-in-the-loop review by human experts. The experts consist of three PhD-level researchers with extensive cooking experience. Based on this process, actions are organized into a two-level hierarchy, whereas objects follow a three-level hierarchical structure. The resulting taxonomy is shown in Figure 6 (a) and (b). Building on this taxonomy, we then design a complementary set of object state change scenarios.

Word Distribution in OSCBench. We visualize the word distribution of all prompts in OSCBench using a word cloud, as shown in Figure 6 (c). This provides an intuitive overview of the dominant concepts and highlights the diversity of objects and actions represented in the benchmark. We further summarize the number of prompts and evaluation dimensions across different T2V benchmarks in Table 5. As shown, OSCBench explicitly emphasizes object state change, complementing existing benchmarks that primarily focus on semantic adherence or physical plausibility.

Benchmarks	#Prompt	SA	PQ	PC	OSC
VBench (Huang et al., 2024)	1362	✓	✓		
EvalCrafter (Liu et al., 2024)	700	✓	✓		
T2V-CompBench (Sun et al., 2025)	1400	✓			
VideoPhy (Bansal et al., 2024)	688	✓		✓	
PhyGenBench (Meng et al., 2024)	160	✓		✓	
PhyWorldBench (Gu et al., 2025)	1050	✓		✓	
OSCBench (ours)	1120	✓	✓		✓

Table 5: Number of prompts and evaluation dimensions in different T2V generation benchmarks. We abbreviate semantic adherence (SA), perceptual quality (PQ), physical commonsense (PC), and object state change (OSC).

B Video Generation Setting

We generate videos for all prompts in our benchmark for each open-source T2V model. For each proprietary T2V model, we generate videos for the selected 140 prompt used in human evaluation. We follow the official and default implementations of T2V models in evaluation. Details of the video generation setting of T2V models, including resolution, total frames, frames per second (FPS), and duration are presented in Table 6. For each generated video, we uniformly sample 20 frames for MLLM-based evaluation. For ViCLIP-based semantic similarity measurement, we uniformly sample 8 frames per video to align with the model architecture.

Models	Resolution	Frames	FPS	Duration (s)
Open-Sora-2.0	768×768	129	25	5
HunyuanVideo	1280×720	129	25	5
HunyuanVideo-1.5	1280×720	121	24	5
Wan-2.2	1280×720	81	16	5
Kling-2.5-Turbo	1920×1080	121	24	5
Veo-3.1-Fast	1280×720	144	24	6

Table 6: Generation settings of T2V models in terms of resolution, total frames, FPS, and duration.

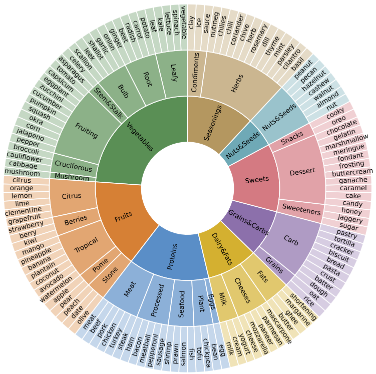
C Evaluation Details

In this section, we provide additional details about our evaluation protocol. We first describe the scoring criteria and then report the results obtained from the MLLM-based evaluation, followed by an illustrative example of how the MLLMs reason and assign scores.

Scoring Criteria. We adopt a hybrid evaluation protocol that combines human user study with automated MLLM-based evaluation. Across these two evaluation modes, we design a comprehensive set of evaluation dimensions covering semantic adherence, OSC performance, scene alignment, and perceptual quality. For each dimension, we provide detailed scoring criteria for both human evaluator and MLLM-based evaluation. The instructions and scoring rubrics in the human user study interface are shown in Figure 7, according to which human evaluators are asked to rate each video on a scale from 1 to 5. For MLLM-based evaluation, the prompts we use are presented in Table 7.

MLLM-based Evaluation Results. We report the additional evaluation results of Qwen3-VL-30B and GPT-5-mini in Table 8 and Table 9, respectively. Although their correlations with human evaluation are not particularly high, the OSC accuracy in both models is lower than their scores on subject

(a) Action taxonomy.



(b) Object taxonomy.

(c) Word cloud of OSCBench.

Figure 6: Data abstraction results and word cloud in OSCBench.

and object semantic adherence. This suggests that Qwen3-VL-30B and GPT-5-mini can capture part of the difficulty associated with object state change.

To further illustrate how MLLMs interpret the videos, Figure 8 presents an example of MLLM evaluation on a generated video. Although the video appears visually appealing, human evaluators identify an OSC error: juice is dripping from the lemon, yet the lemon itself shows no visible squeezing or deformation. For this video, we observe that Qwen3-VL-30B and GPT-5-mini assign perfect scores, indicating that they fail to detect the fine-grained issues in the lemon's state change. GPT-5.2, in contrast, is able to detect the OSC error and provides reasonable supporting evidence, noting that "the lemon largely remains undeformed". This suggests that more advanced MLLMs can to handle state-change reasoning and identify inconsistencies between visual appearance and expected physical outcomes. Although GPT-5.2 w/o CoT also assigns a relatively low score to OSC accuracy, it still gives action alignment and OSC accuracy the same score, which shows weaker consistency with human judgments. From the human perspective, the error in action alignment is minor, whereas the error in OSC is a major one. This indicates that using CoT to plan a reasoning route, in which the model first follows the grading guidelines to collect explicit evidence and then assigns scores, encourages more careful evaluation and results in scores that better align with human judgments.

D Examples of Different OSC Scenarios

To provide a more intuitive view of T2V performance on object state change in OSCBench, we present examples of generated videos in regular, novel, and compositional OSC scenarios in Figures 9, Figure 10, and Figure 11, respectively.

In the regular OSC scenario shown in Figure 9, all models generate videos in which the subject (chef), action (slicing), object (leek), and scene (street food stand) are rendered well. However, clear errors emerge in the object state change. For example, in videos generated by Open-Sora-2.0, HunyuanVideo, HunyuanVideo-1.5, and Kling-2.5-Turbo, the leek is not actually sliced into pieces. Although videos produced by Wan-2.2 and Veo-3.1-Fast exhibit correct object state changes, the state change consistency in the later frames is remains limited.

In the novel OSC scenario shown in Figure 10, models can roughly understand the peeling action, but noticeable issues remain. The hand details in Open-Sora-2.0 are blurred. In HunyuanVideo, HunyuanVideo-1.5, and Wan-2.2, the object being peeled is incorrect, with Wan-2.2 generating olives, which are more commonly associated with the peeling action. Kling-2.5-Turbo exhibits state change consistency issues, where two berries gradually collapse into one. Veo-3.1-Fast produces an OSC that is close to correct, but the object still shows jitter and artifacts in the last frame. Overall, the novel scenario remains challenging for current T2V models.

In the compositional OSC scenario shown in Figure 11, Open-Sora-2.0, HunyuanVideo, Wan-2.2, and Kling-2.5-Turbo execute only one of the required actions. Although Veo-3.1-Fast successfully completes the compositional actions, the consistency of the object states is poor. For example, a spatula suddenly appears in frame 4, and the ham disappears in frame 5, revealing noticeable artificial artifacts. These results suggest that compositional OSC remains difficult for current T2V models.

Suppose you are an expert in judging and evaluating the quality of AI-generated videos. Such videos may exhibit anomalies such as unnatural object appearance or disappearance, physically implausible state changes, and temporal inconsistencies across frames. They may also contain visual artifacts or unnatural textures. You are given 20 frames evenly sampled from a 5-second AI-generated video.
Video Prompt:
"A chef with a white apron is slicing leek at a street food stand."

Your Task:
Analyze these frames chronologically and evaluate the video using the following criteria.
{Criteria}

Instructions:
- Evaluate each criterion INDEPENDENTLY.
- For each criterion, first identify the relevant factual evidence from the frames, then assign a score.

Output Format:
Return the result strictly in JSON format.
{
"Subject Alignment": "evidence": "...", "score": [1-5],
"Object Alignment": "evidence": "...", "score": [1-5],
"Action Alignment": "evidence": "...", "score": [1-5],
"OSC Accuracy": "evidence": "...", "score": [1-5],
"OSC Consistency": "evidence": "...", "score": [1-5],
"Scene Alignment": "evidence": "...", "score": [1-5],
"Realism": "evidence": "...", "score": [1-5],
"Aesthetics": "evidence": "...", "score": [1-5]
}

Table 7: Prompt for MLLM to generate the evidence and score for each sampled video. Criteria are the same as those used in the human-evaluation interface.

Model	Semantic Adherence			Object State Change		Scene Alignment	Perceptual Quality	
	Subject	Object	Action	Accuracy	Consistancy		Realism	Aesthetics
Open-source models								
Open-Sora-2.0 (Peng et al., 2025)	0.978	0.932	0.718	0.656	0.722	0.980	0.796	0.816
HunyuanVideo (Kong et al., 2024)	0.966	0.954	0.682	0.648	0.752	0.988	0.846	0.858
HunyuanVideo-1.5 (Team, 2025)	0.986	**0.958**	**0.780**	**0.756**	0.808	0.990	0.864	0.864
Wan-2.2 (Wan et al., 2025)	**0.992**	0.952	0.738	0.712	**0.820**	**0.992**	**0.882**	**0.910**
Proprietary models								
Kling-2.5-Turbo (KlingAI, 2025)	**0.998**	0.972	0.934	0.882	0.906	**0.998**	0.950	0.952
Veo-3.1-Fast (Google DeepMind, 2025b)	0.996	**0.988**	**0.970**	**0.954**	**0.968**	0.996	**0.994**	**0.978**

Table 8: Qwan3-VL-30B-based evaluation results of different T2V models across multiple evaluation dimensions.

Model	Semantic Adherence			Object State Change		Scene Alignment	Perceptual Quality	
	Subject	Object	Action	Accuracy	Consistancy		Realism	Aesthetics
Open-source models								
Open-Sora-2.0 (Peng et al., 2025)	0.918	0.838	0.712	0.794	0.814	0.930	0.722	0.774
HunyuanVideo (Kong et al., 2024)	0.960	0.942	0.792	0.718	0.956	0.996	0.908	0.900
HunyuanVideo-1.5 (Team, 2025)	0.934	0.914	0.794	0.582	0.788	0.962	0.834	0.788
Wan-2.2 (Wan et al., 2025)	**0.946**	**0.944**	**0.750**	**0.744**	**0.970**	**0.966**	**0.930**	**0.946**
Proprietary models								
Kling-2.5-Turbo (KlingAI, 2025)	**0.996**	0.982	0.976	0.906	**0.998**	0.996	**0.980**	**1.000**
Veo-3.1-Fast (Google DeepMind, 2025b)	0.992	**0.994**	**0.994**	**0.966**	**0.998**	**0.998**	0.958	0.984

Table 9: GPT-5-mini-based evaluation results of different T2V models across multiple evaluation dimensions.

Task Instructions

Please use the same evaluation criteria to score all videos and follow the definitions below.
1. Read the **prompt** and watch the **video** from start to finish.
2. Evaluate each criterion **independently**.
   Example: when scoring Action Alignment, focus only on the correctness of the action, independent of the object and other attr ibutes.
3. Use the **1–5 scale** consistently across all criteria and all videos.

Evaluation Criteria

1. Semantic Adherence

1a Subject Alignment

Is the subject present and correct (i.e., the main actor, e.g., a person or a hand)?
(Please focus only on the subject. Please select "NA" if the prompt does not specify a subject.)

1. **Very poor**: Subject is absent or replaced by something entirely unrelated.
2. **Poor**: Subject is present but does not match the expected category.
3. **Fair**: Subject is of the correct category but exhibits major attribute errors.
4. **Good**: Subject is correct and well-rendered, with only minor attribute errors.
5. **Excellent**: Subject perfectly matches the prompt in category, form, and attributes.

1b Manipulated Object Alignment

Is the manipulated object present and correct (e.g., carrots or tomatoes)?

1. **Very poor**: Manipulated object is absent, or a completely different object is present.
2. **Poor**: Manipulated object is of the wrong category or is severely distorted.
3. **Fair**: Manipulated object is of the correct category but shows major visual inaccuracies.
4. **Good**: Manipulated object is correct and realistic, with only minor visual inaccuracies.
5. **Excellent**: Manipulated object is realistic and provides a perfect visual match.

1c Action Alignment

Does the performed action match the action in the prompt (e.g., slicing or roasting)?

1. **Very poor**: A fundamentally different action is performed.
2. **Poor**: The intended action is recognizable but executed in a physically incorrect way.
3. **Fair**: The correct action is performed but with clear physical or logical flaws.
4. **Good**: Action is performed correctly, but motion appears slightly unnatural.
5. **Excellent**: Action is executed in a physically plausible, natural manner.

2. State Change Performance

2a Object State Change Accuracy

Is the object state change correct and as expected (e.g., an apple changing from whole to slices)?

1. **Very poor**: Object state change is illogical or unrelated to the action.
2. **Poor**: Object state change clearly does not match the expected outcome.
3. **Fair**: Object state change is partially correct, but major inaccuracies remain.
4. **Good**: Object state change is generally correct, with minor issues.
5. **Excellent**: Object state change is accurate and matches the expected outcome exactly.

2b Object Change Continuity & Consistency

Is the object state change continuous and natural, without any unnatural object appearances or disappearances?

1. **Very poor**: State change is highly discontinuous, with obvious jumps or objects suddenly appearing/disappearing.
2. **Poor**: State change is discontinuous or has noticeable object appearances/disappearances.
3. **Fair**: State change is mostly continuous but includes small jumps or object inconsistencies.
4. **Good**: State change is continuous and natural, with only minimal, non-disruptive inconsistencies.
5. **Excellent**: State change is smooth and continuous, with no unnatural object appearances/disappearances.

3.SCENE

3a Scene Alignment

Does the background and environment match the prompt (e.g., a kitchen or a market)?
(Please focus only on the scene and environment. Please select "NA" if the prompt does not specify a scene.)

1. **Very poor:** Scene directly contradicts the prompt.
2. **Poor:** Scene is generic or ambiguous and lacks required details.
3. **Fair:** Scene partially matches the prompt but contains notable attribute inaccuracies.
4. **Good:** Scene contains correct elements with only minor attribute inaccuracies.
5. **Excellent:** Scene is a detailed and accurate match to the prompt's setting.

4. Perceptual Quality

4a Realism

Does this video look like a real-world video?

1. **Very poor**: Video looks artificial, distorted, or obviously fake.
2. **Poor**: Many visual artifacts; motion, lighting, or textures do not resemble real footage.
3. **Fair**: Some elements look real, but noticeable artifacts reduce overall realism.
4. **Good**: Video appears close to real with only minor visual imperfections.
5. Excellent: Video looks convincingly real with natural motion, lighting, and textures.

4b Aesthetic

Is the video visually appealing? Are the colors harmonious and is the content rich?

1. **Very poor**: Video is visually unappealing, with distracting colors or dull/empty content.
2. **Poor**: Some attempt at aesthetics, but colors clash or the content feels sparse.
3. **Fair**: Overall visually fine, with moderate harmony and adequate content richness.
4. **Good**: Visually appealing, with harmonious colors and rich, engaging content.
5. **Excellent**: Highly pleasing visuals, strong color harmony, and rich, well-composed content throughout.

Figure 7: Task instructions and evaluation criteria in the human evaluation interface.

Prompt: A woman in casual clothes is squeezing lemon on the grass.

Evaluator	Subject Alignment	Object Alignment	Action Alignment	OSC Accuracy	OSC Consistency	Scene Alignment	Realism	Aesthetics
Human	4.7	4.7	4.3	3.3	4.0	4.7	3.7	4.7
Qwen3-VL-30B	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
GPT-5-mini	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0
GPT-5.2 w/o CoT	4.0	4.0	3.0	3.0	3.0	5.0	4.0	4.0
GPT-5.2	5.0	4.0	4.0	3.0	3.0	5.0	4.0	4.0

Evidence from GPT-5.2 Evaluation

Subject Alignment	Object Alignment	Action Alignment	OSC Accuracy
Across all frames, a woman is clearly visible sitting on grass in an outdoor field, wearing casual clothing.	A lemon remains in her right hand while liquid drips down, and additional lemons sit on a plate near the bottom of the frame.	The woman holds a lemon appears to squeeze it. However, the hand compression of the lemon is minimal and the left hand is not clearly assisting the squeeze.	Expected state change would be lemon being squeezed/partially deformed or cut and juice expressed. The video shows juice dripping, but the lemon largely remains undeformed and not visibly cut is shown.

OSC Consistency	Scene Alignment	Realism	Aesthetics
Juice stream persists fairly consistently frame-to-frame, but there are continuity issues: the lemon stays visually whole while producing juice, and in later frames an additional lemon appears in her left hand without a clearly shown pickup motion.	Background consistently depicts an outdoor grassy area with plants a bright sky, matching the prompt's requirement	Overall looks close to real footage with natural lighting and depth of field. Minor AI cues include too-smooth skin, somewhat frozen motion, and the physically questionable continuous juice from an apparently whole lemon.	Pleasant composition and color harmony (greens of grass, blue sky). Outdoor scenery is rich and well-lit; subject is visually appealing.

Figure 8: Example of MLLM evaluation on a generated video. Original human evaluation scores averaged over three evaluators are provided for reference. The evidence is generated by GPT-5.2 when scoring with CoT.

Regular Scenario: A chef with a white apron is slicing leek at a street food stand.

Figure 9: Sampled videos of different models in regular OSC scenario.

Figure 10: Sampled videos of different models in novel OSC scenario.

Figure 11: Sampled videos of different models in compositional OSC scenario.