From Evaluation to Enhancement: Benchmarking and Improving Think-with-Video Reasoning for Video Generative Models

Meng Luo[1,*,†], Yicheng Liu[2,*,†], Jiahao Wang[3,‡], Yuanxing Zhang[3], Xin Tao[3], Pengfei Wan[3], Kun Gai[3], and Hao Fei[4,‡]

[1] School of Computing, National University of Singapore, Singapore, Singapore
[2] Department of Computer Science and Technology, Nanjing University, Nanjing, China
[3] Kling Team, Kuaishou Technology, Beijing, China
[4] Department of Computer Science, University of Oxford, Oxford, United Kingdom

Abstract. Video generation has advanced to produce visually compelling and temporally coherent results. Yet, whether these models can genuinely think with video—executing symbolic rules, respecting physical laws, and pursuing intentional goals—remains an open question. Existing benchmarks only partially address this, often conflating visual quality with cognitive correctness. We introduce `VWG-Bench` (**V**ideo **W**orld **G**eneralist Benchmark), a comprehensive benchmark spanning 9 reasoning dimensions and 38 fine-grained tasks. To enable precise diagnosis, we design a three-level VLM-as-Judge protocol that independently assesses video-level fluency, task-level rule adherence, and sample-level goal realization. Evaluations of leading models reveal a striking gap: while models achieve strong rendering scores, they consistently fail on logic-heavy and rule-constrained tasks. To address this, we propose `Vid-PRE` (**Vid**eo **P**rompt **R**easoner and **E**nhancer), a model-agnostic prompt rewriter that offloads the cognitive burden of reasoning to a dedicated VLM. Trained via reinforcement learning with purely text-based rewards, Vid-PRE produces concise, constraint-aware prompts without the instability of video-level reward signals. Experiments show that Vid-PRE yields substantial reasoning improvements across multiple generators without architectural modifications. Together, `VWG-Bench` and `Vid-PRE` offer a rigorous diagnostic lens and a scalable path toward true think-with-video capabilities. All data and code are publicly available at `https://huggingface.co/datasets/KlingTeam/VWG-Bench`.

Keywords: Video Generation · Evaluation Benchmark · Prompt Enhancement · Reinforcement Learning

---

* Equal contribution.
† Work done during an internship at Kling Team, Kuaishou Technology.
‡ Corresponding authors.

1 Introduction

Reasoning through continuous spatio-temporal signals is a hallmark of intelligence that text and static images cannot fully replicate [7, 22, 28, 45]. "Thinking with Text" established Chain-of-Thought (CoT) [27, 30, 37, 42] prompting as a foundational paradigm for language reasoning; "Thinking with Images" extended this to visual chains in Vision-Language Models [23, 33, 44]. A natural next frontier is *Think with Video*: reasoning expressed through continuous video frames that encode physical dynamics, object interactions, and long-horizon goal execution in a unified temporal substrate. Recent advances, including Veo 3's chain-of-frames, Sora-2's visual problem solving [13, 26, 34], and emergent spatial planning in open-source models, suggest that video generation is crossing a threshold from photorealistic rendering toward genuine world simulation [6, 24, 36]. Yet a fundamental question remains unresolved: do these models truly understand the rules they appear to follow, or are they exploiting surface statistics to produce plausible-looking sequences?

Answering this question requires benchmarks that probe reasoning rather than aesthetics. Existing efforts have made important strides: TiViBench [4] introduces hierarchical evaluation across four structural and logical dimensions; V-ReasonBench [29] provides reproducible, answer-verifiable tasks grounded in spatial and physical cognition; VR-Bench [40] systematically measures spatial planning through maze-solving; RULER-Bench [16] targets cognitive-rule adherence across six categories; RISE-Video [25] probes implicit world-rule understanding across eight knowledge domains; and MMGR [3] covers abstract and embodied reasoning. Despite this progress, the evaluation landscape remains fragmented. Each benchmark covers only a subset of the reasoning spectrum. Moreover, existing protocols typically collapse evaluation into a single binary or holistic score, masking where and why models fail, specifically whether at the level of visual fluency, task-level rule tracking, or precise goal realization.

To close both gaps, we introduce **`VWG-Bench`**, a comprehensive and hierarchical benchmark designed to stress-test think-with-video reasoning across the full spectrum of world simulation. As illustrated in Figure 1, `VWG-Bench` covers **9** core reasoning dimensions and **38** fine-grained tasks. Specifically, these dimensions systematically evaluate: (1) Symbolic Reasoning & Logic Puzzles, (2) Graphical Reasoning & Pattern Matching, (3) Embodied & Experiential Reasoning, (4) GUI & Digital Interaction, (5) Hypothetical & Counterfactual Reasoning, (6) Spatial Geometry & Logic, (7) Multiscale Temporal Reasoning, (8) Natural Laws & World Simulation, and (9) Social Understanding & Commonsense. To construct evaluation samples at scale, we develop a highly automated data pipeline that orchestrates image generation, consistency filtering, and intent formulation, enabling scalable production of high-quality instances.

Beyond scale and coverage, `VWG-Bench` introduces a three-level VLM-as-Judge evaluation framework with a fine-grained 1–5 scoring scale. Unlike existing holistic scoring that conflates rendering quality with cognitive correctness, our protocol explicitly decouples three orthogonal facets of performance: *Video-level*, assessing overall visual fluency and temporal coherence; *Task-level*, measuring

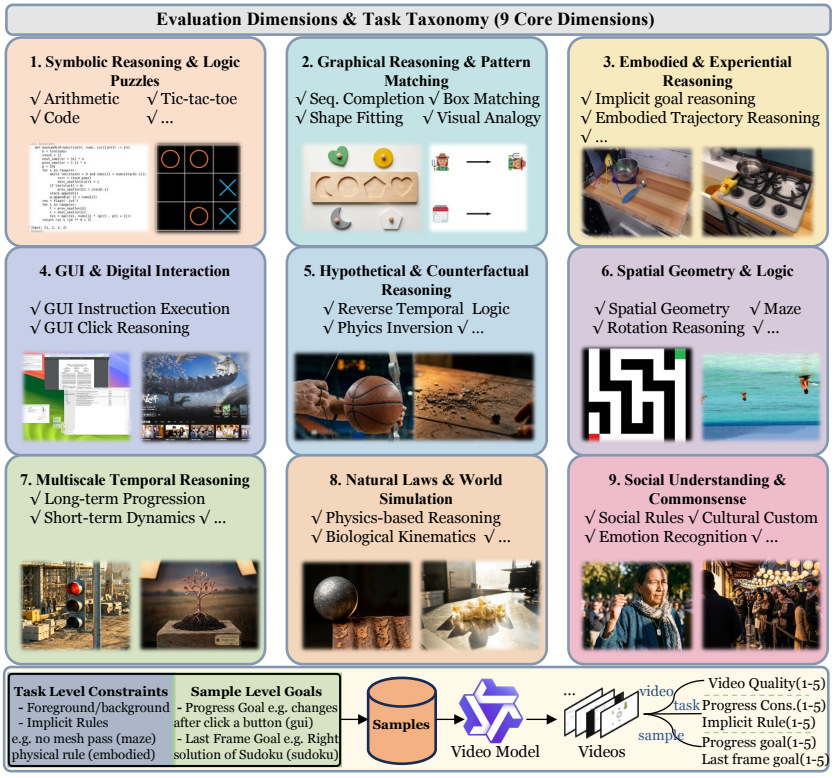

Fig. 1: Comprehensive overview of `VWG-Bench`'s taxonomy, automated data construction, and structured evaluation framework.

progress consistency and adherence to implicit task rules across frames; and *Sample-level*, evaluating the precise realization of intermediate milestones and final objectives. This decomposition enables fine-grained, interpretable diagnostics that identify failure modes at each level of the reasoning hierarchy. Extensive evaluations of leading commercial and open-source models on VWG-Bench expose a consistent pattern: models achieve strong Video-level rendering scores but suffer severe degradation at the Task- and Sample-level on compositional, rule-constrained, and logic-heavy scenarios. This gap is not primarily a failure of visual generation; it is a failure of reasoning through prompts: given only a natural-language description, models lack the structured representation of task rules, success criteria, and constraint hierarchies needed to guide generation toward correct reasoning trajectories.

To directly address this limitation, we propose **`Vid-PRE`**, a model-agnostic prompt rewriter that decouples the reasoning stage from the rendering stage. `Vid-PRE` offloads the cognitive burden to a dedicated VLM, which infers task rules and success criteria from the image context and distills them into a concise,

constraint-aware prompt for the downstream video generator. `Vid-PRE` is trained via Supervised Fine-tuning (SFT) on rejection-sampled reasoning tuples and further optimized with Group Relative Policy Optimization (GRPO) [32] using a meticulously designed purely text-based rewards system, keeping it agnostic to any specific generator architecture. Extensive experiments confirm substantial improvements across multiple generators without any architectural modification.

In summary, our contributions are threefold:

- **A Unified Think-with-Video Benchmark.** `VWG-Bench` is the first benchmark to jointly cover all nine core dimensions of think-with-video world simulation across 38 tasks, constructed via a fully automated data pipeline for scalable and reproducible scenario generation.
- **A Fine-Grained, Hierarchical Evaluation Protocol.** Our three-level VLM-as-Judge framework with a 1–5 scoring scale diagnoses Video-level fluency, Task-level rule adherence, and Sample-level goal realization, revealing failure modes that holistic or binary evaluation cannot.
- **A Universal and Plug-and-Play Prompt Reasoner and Enhancer.** `Vid-PRE` is a model-agnostic rewriter trained via SFT and GRPO with text-based rewards that significantly boosts reasoning accuracy across multiple video generators, without any modification to their architectures.

2 `VWG-Bench`: Video World Generalist Benchmark

2.1 Design Principles

VWG-Bench is designed to comprehensively evaluate the reasoning capabilities of video generation models. Compared with prior reasoning benchmarks for video generation, VWG-Bench aims to support both broader task coverage and a more systematic evaluation framework.

For task coverage, as shown in Table 1, our benchmark spans nine major reasoning dimensions and 38 task categories, fully covering symbolic and graphical reasoning, embodied and interactive tasks, as well as spatial–temporal reasoning, physical world simulation, and social commonsense understanding. To balance evaluation comprehensiveness with practical cost, we sample only 10 high-quality instances from each task category, significantly reducing the computational and annotation overhead required for evaluation.

For the evaluation framework, as illustrated in Figure 1, we establish a three-level evaluation protocol. The video-level performs coarse-grained quality control to ensure the generated video is a valid visual medium. The task-level verifies whether the model respects task-specific constraints and logical rules throughout the video generation process. Finally, the sample-level assesses whether the final outcome satisfies the intended task objective. Compared to prior benchmarks that typically involve only one or two evaluation dimensions, our framework provides a more comprehensive and structured assessment of reasoning performance.

2.2 Evaluation Dimensions and Task Taxonomy

Here we introduce the nine evaluation dimensions and the 38 task categories covered by VWG-Bench.

VWG-Bench organizes tasks into nine reasoning dimensions spanning symbolic logic, visual pattern induction, embodied interaction, digital interfaces, counterfactual reasoning, spatial geometry, temporal evolution, world simulation, and social understanding. **Symbolic Reasoning & Logic Puzzles** evaluate the execution of formal rules and algorithms, such as "Arithmetic", "Coding", and "Sudoku". **Graphical Reasoning & Pattern Matching** focuses on discovering visual correspondences and completing structured patterns, such as "Box Matching", "Color Connecting", and "Visual Analogy". **Embodied & Experiential Reasoning** examines reasoning in physical interaction scenarios, including tasks like "Embodied Action Execution", "Embodied Trajectory Reasoning", and "Implicit Goal Reasoning". **GUI & Digital Interaction** evaluates reasoning within graphical user interfaces, such as "GUI Instruction Execution" and "GUI Click Reasoning". **Hypothetical & Counterfactual Reasoning** probes reasoning under altered or fictional premises, for example "Reverse Temporal Logic", "Physics Inversion", and "Causal Intervention". **Spatial Geometry & Logic** tests geometric consistency and spatial transformations, including tasks such as "Maze", "Rotation Reasoning", and "Spatial Geometry". **Multiscale Temporal Reasoning** evaluates the ability to maintain coherent dynamics across different time horizons, such as "Short-term Temporal Dynamics" and "Long-term Temporal Progression". **Natural Laws & World Simulation** measures whether models follow real-world physical or biological principles, for example "Physics-based Reasoning", "Biological Kinematics", and "Physical Commonsense Reasoning". **Social Understanding & Commonsense** focuses on human-centric semantics and social behavior modeling, including tasks like "Social Rules", "Cultural Customs", and "Emotion Recognition".

2.3 Data Construction Pipeline

Image-Pair Initialization. To collect prompt-image pairs across such a diverse set of task categories, we adopt three complementary strategies: (1) identifying suitable open-source datasets and sampling high-quality categories from them (e.g., GUI and embodied tasks); (2) generating samples through programmatic algorithms (e.g., arithmetic and Sudoku tasks); or (3) automatically generating images via Text-to-Image (T2I) models using predefined prompt templates and dynamic variable pools to ensure broad scenario diversity. To guarantee visual integrity, we apply a strict visual filtering mechanism to T2I outputs. A VLM acts as an impartial judge to verify whether the generated image faithfully adheres to the T2I prompt without missing elements or visual hallucinations. If the image fails this check, the pipeline either regenerates the image or discards it entirely.

Multi-Granularity Evaluation Annotation. To support our fine-grained, three-level evaluation framework, the pipeline automatically annotates diagnos-

Table 1: Comparison of `VWG-Bench` with representative video-generation reasoning benchmarks. **Sym**: Symbolic & Logic, **Gph**: Graphical & Pattern, **Emb**: Embodied & Experiential, **GUI**: GUI & Digital Interaction, **Cft**: Hypothetical & Counterfactual, **Spa**: Spatial Geometry, **Tmp**: Multiscale Temporal, **Phy**: Natural Laws & Physics, **Soc**: Social & Commonsense). **Auto** = whether an automated data-construction pipeline is provided. ✓ = fully covered; ○ = partially covered; ✗ = not covered.

	Scale			Reasoning Coverage									Evaluation	
Benchmark	**Dims**	**Tasks**	**Samples**	**Sym**	**Gph**	**Emb**	**GUI**	**Cft**	**Spa**	**Tmp**	**Phy**	**Soc**	**Levels**	**Auto**
MME-CoF [15]	/	12	59	✗	✗	○	✓	✗	✓	○	✓	✗	1	✗
VideoThinkBench [34]	2	5	4,149	○	✓	✗	✗	✗	○	✗	○	✗	1	✗
TiViBench [4]	4	24	595	✓	✓	○	✗	✗	✓	✗	✗	✗	1	✗
VR-Bench [40]	1	5	7,920	✗	✗	✗	✗	✗	✓	✗	✗	✗	1	✗
V-ReasonBench [29]	4	12	327	✓	✓	✗	✗	✗	✓	✗	✓	✗	1	✗
RULER-Bench [16]	6	40	622	✓	✗	✗	✗	✗	✓	○	✓	✗	1	✗
MMGR [3]	3	10	1,853	✓	✗	✓	✗	✗	✓	✗	✓	✗	1	✗
RISE-Video [25]	8	29	467	✓	○	✓	✗	✗	✓	✓	○	✓	1	✓
VWG-Bench (Ours)	**9**	**38**	**380**	✓	✓	✓	✓	✓	✓	✓	✓	✓	**3**	✓

tic criteria that form the core ground truth for VWG-Bench. For task-level constraints, we define rule dictionaries that delineate the expected behaviors of the foreground (e.g., allowed state changes for manipulated objects), the background (e.g., strict static invariance for unmanipulated areas), and implicit rules (e.g., forbidding mesh interpenetration during object stacking). For sample-level goals, the VLM simulates the outcome of the action to generate linguistic descriptions of the expected final state (i.e., Last-Frame Goal, $\mathcal{G}_{\text{last}}$). For tasks with complex temporal dynamics, intermediate milestones (i.e., Progress Goal, $\mathcal{G}_{\text{prog}}$) are additionally annotated.

2.4 Evaluation Framework and Protocols

Our evaluation framework is designed to systematically assess the reasoning capabilities of video generation models, prioritizing logical coherence and physical adherence over mere visual fidelity. To mitigate evaluation hallucinations, we employ a VLM-as-Judge paradigm, where the judge strictly cross-references generated frames against structured annotations. All metrics across our three-level framework are scored on a 1–5 integer scale (according to the degree of consistency with the corresponding annotations).

Video-Level Assessment (Coarse-Grained Quality Control). Before probing reasoning, we must ensure the generated video is a valid visual medium. This foundational filter evaluates Video Quality by assessing holistic visual fidelity and temporal smoothness. The VLM judge explicitly penalizes technical flaws that obscure the video content—such as severe camera jitter, temporal incoherence, unnatural flickering, or excessive static frames—ensuring that poor baseline rendering does not confound the subsequent reasoning evaluation.

Task-Level Assessment (Fine-Grained Consistency & Constraints). This level leverages rule dictionaries to verify whether the model respects spatial

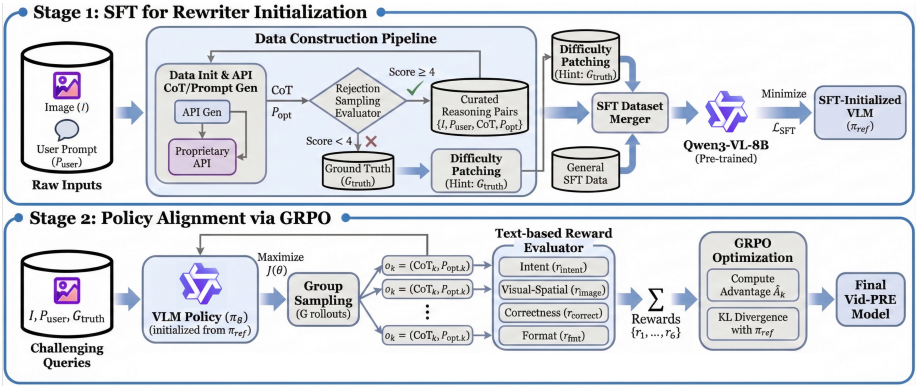

Fig. 2: An overview of the training framework for `Vid-PRE`.

boundaries and logical constraints throughout the generated video. Specifically, Progress Consistency scrutinizes the temporal coherence of dynamic elements by checking whether the foreground evolves logically while penalizing unprompted changes or morphological collapse in the background. Implicit Rule Following assesses adherence to physical laws and hidden boundaries inherent to the task, strictly penalizing violations such as objects interpenetrating solid meshes, reversing gravity, or rigid bodies undergoing unreasonable soft deformations.

Sample-Level Assessment (Goal Realization). The final level evaluates the semantic alignment between the temporal states of the video and the user's explicit instructions. Progress Goal evaluates necessary intermediate milestones; for complex temporal progressions, the judge inspects intermediate keyframes to ensure the "Start → Intermediate → End" sequence is correctly chained without skipping or reversing logical steps. Last-Frame Goal assesses ultimate task completion by comparing the terminal state of the generated video against the annotated goal description to determine whether the objective was achieved.

3 `Vid-PRE`: A Universal Prompt Reasoner and Enhancer

Current Image-to-Video (I2V) generation models excel at visual rendering but struggle with complex logical derivation and constraint satisfaction. To address this, we introduce Vid-PRE, a model-agnostic prompt rewriting module. The core philosophy of Vid-PRE is a strict division of labor: it offloads the cognitive burden of spatio-temporal reasoning to a dedicated VLM, allowing the downstream video generator to focus exclusively on rendering based on highly explicit and executable instructions.

3.1 Problem Formulation

We formulate video prompt rewriting as a conditional text generation problem. Let $\mathcal{I}$ denote the first-frame image and $\mathcal{P}_{\text{user}}$ the user instruction. A naive for-

mulation would learn a direct mapping $(\mathcal{I}, \mathcal{P}_{\text{user}}) \rightarrow \mathcal{P}_{\text{opt}}$, but this often leads to prompts that lack explicit logical grounding. To encourage structured reasoning during rewriting, we introduce an intermediate representation CoT that captures the reasoning process linking the visual context and the final executable prompt. The objective of Vid-PRE is therefore to model the joint distribution $P(\text{CoT}, \mathcal{P}_{\text{opt}} \mid \mathcal{I}, \mathcal{P}_{\text{user}})$. Let $x = (\mathcal{I}, \mathcal{P}_{\text{user}})$ denote the input and $y = [\text{CoT}; \mathcal{P}_{\text{opt}}]$ the concatenated output sequence. The generation process follows a standard autoregressive factorization:

$$P(y \mid x) = \prod_{t=1}^{|y|} P(y_t \mid y_{<t}, x).$$

This formulation encourages the model to first articulate the reasoning required to interpret the visual scene and constraints, and then produce a concise executable prompt $\mathcal{P}_{\text{opt}}$.

3.2 Stage 1: Supervised Fine-Tuning with Reasoning Data

The first stage initializes the VLM with the structural capability to follow I2V formatting conventions and execute rigorous reasoning. We construct a high-quality dataset of approximately 20K tuples $\{\mathcal{I}, \mathcal{P}_{\text{user}}, \text{CoT}, \mathcal{P}_{\text{opt}}\}$ through an automated rejection sampling paradigm using advanced proprietary APIs. As illustrated in the upper part of Figure 2, the data construction process follows four steps:

Step1: Data Initialization. For each given pair $(\mathcal{I}, \mathcal{P}_{\text{user}})$, we computationally define the rigorous ground truth ($\mathcal{G}_{\text{truth}}$, *e.g.*, the expected last-frame state or progress goal).

Step2: CoT and Prompt Generation. The API analyzes the initial image, clarifies the user's intent, and identifies critical constraints to produce a natural-language CoT. It then generates a concise prompt $\mathcal{P}_{\text{opt}}$ suitable for video generation.

Step3: Rejection Sampling. To ensure the reliability of generated samples, we introduce an independent evaluator to assess the quality of each candidate $\mathcal{P}_{\text{opt}}$. The evaluator has access to the input pair $(\mathcal{I}, \mathcal{P}_{\text{user}})$ as well as the corresponding ground truth $\mathcal{G}_{\text{truth}}$, and evaluates the output from two perspectives. First, it verifies visual consistency, checking whether the prompt introduces hallucinated objects, incorrect spatial relations, or directional misunderstandings with respect to the image. Second, it assesses whether the optimized prompt logically leads to the target state defined by $\mathcal{G}_{\text{truth}}$. The evaluator then assigns a comprehensive score on a 1–5 scale. Only samples with scores ≥ 4 are retained together with their corresponding CoT. This rejection sampling strategy ensures that the reasoning process and the resulting prompt remain logically coherent and consistent.

Step4: Difficulty Patching. For exceptionally hard task categories where the API fails to deduce a valid logical chain under zero-shot conditions, we

relax the constraint by providing $\mathcal{G}_{\text{truth}}$ as a hint, allowing it to generate critical logical patches for the CoT. The patched outputs are then re-evaluated through the same rejection sampling pipeline in Step (3) to maintain quality control.

These curated reasoning pairs are mixed with general SFT data—to preserve the base conversational capabilities of the VLM—and used to fine-tune a pre-trained Qwen3-VL-8B [1] model. The optimization minimizes the standard negative log-likelihood loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{\mathcal{D}_{\text{SFT}}} \left[\log P_\theta(\text{CoT}, \mathcal{P}_{\text{opt}} \mid \mathcal{I}, \mathcal{P}_{\text{user}}) \right]. \tag{1}$$

3.3 Stage 2: Policy Alignment via GRPO

While SFT provides a strong logical initialization, performance on complex, long-horizon reasoning tasks remains suboptimal. As illustrated in the bottom part of Figure 2, to further unlock reasoning potential, we apply GRPO using 1.5K specialized prompts collected across 10 challenging tasks that exhibit the highest error rates after the SFT stage.

A key design decision is our adoption of a purely text-based reward system in lieu of video-generation-based rewards. This choice is motivated by three considerations: (1) **Stability and Efficiency**: Current video models exhibit substantial aesthetic variance, where poor rendering can easily suppress the reward signal of a logically correct prompt. Text-based evaluation avoids this instability and the considerable computational cost of rendering videos during RL training. (2) **Generator Agnosticism**: By decoupling from any specific video generator, Vid-PRE acquires no model-specific priors, functioning as a universal plug-and-play module. (3) **Task Nature**: I2V reasoning tasks are tightly bounded by the initial frame and deterministic physical/geometric constraints; thus, optimizing the textual prompt's semantic correctness directly addresses the root cause of reasoning failures.

During training, for each query $x = (\mathcal{I}, \mathcal{P}_{\text{user}}, \mathcal{G}_{\text{truth}})$, the policy π_θ samples a group of G outputs $\{o_1, \ldots, o_G\}$, where $o_k = (\text{CoT}_k, \mathcal{P}_{\text{opt},k})$. Our reward function evaluates each rollout along three dimensions on a 1–5 scale. **Intent Preservation** (r_{intent}) verifies that $\mathcal{P}_{\text{opt},k}$ reasonably extends and refines $\mathcal{P}_{\text{user}}$ without altering the core objective. It ensures key objects and goals are preserved, while minor wording differences (*e.g.*, camera terminology) are not overly penalized. **Visual-Spatial Consistency** (r_{image}) checks whether the spatial relations described in $\mathcal{P}_{\text{opt},k}$ (*e.g.*, "bottom right") accurately reflect the actual content of the image $\mathcal{I}$, strictly penalizing hallucinated visual elements or invalid spatial relationships. **Solution Correctness** (r_{correct}) evaluates whether the explicit steps inferred in the prompt correctly solve the task and reach $\mathcal{G}_{\text{truth}}$, emphasizing critical problem-solving details (*e.g.*, specific paths or action steps) over generic generation constraints. In addition, a format compliance score (r_{format}) ensures outputs adhere to the prescribed structural template. The final reward is a weighted combination: $r_k = w_1 r_{\text{intent}} + w_2 r_{\text{image}} + w_3 r_{\text{correct}} + w_4 r_{\text{format}}$.

GRPO optimizes the policy by maximizing the advantage of outputs with higher relative rewards within the group [9, 10, 39]. The advantage for each rollout is computed as the group-normalized reward: $\hat{A}_k = (r_k - \mu_G)/\sigma_G$, where

μ_G and σ_G are the mean and standard deviation of rewards within the group $\{r_1, \ldots, r_G\}$. Let π_{ref} denote the SFT-initialized policy serving as the fixed reference for KL regularization:

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{\text{RL}},\, o \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{k=1}^{G} \frac{\pi_\theta(o_k \mid x)}{\pi_{\theta_{\text{old}}}(o_k \mid x)} \hat{A}_k - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right]. \quad (2)$$

Through this paradigm, Vid-PRE learns to produce logically precise and physically grounded reasoning chains that translate into robust instructions for downstream video generation.

4 Experiments

4.1 Settings

We evaluate the general reasoning capability of six video generation models on VWG-Bench, including one open-weight model (Wan2.2) and five commercial systems (Wan2.5, Wan2.6-Flash, Kling-2.5-Turbo-Pro, Sora2, and Veo3.1). For each $(\mathcal{I}, \mathcal{P}_{\text{user}})$ pair, we sample three generations and report the averaged score for each evaluation dimension. During evaluation, we extract the last frame of each generated video to assess $\mathcal{G}_{\text{last}}$, while the remaining metrics are evaluated using a frame sampling configuration of 2 fps with a maximum of 16 frames.

We initialize Vid-PRE with the pre-trained Qwen3-VL-8B [1] model, and use Gemini-3.0-Pro as the data generator and the unified evaluator. During the SFT stage, the model is trained on approximately 20K high-quality samples for 3 epochs with a batch size of 128. The maximum image token length is constrained to 1,024 to balance fine-grained visual detail and computational efficiency. In the subsequent GRPO stage, the policy is further optimized using 1.5K high-difficulty samples for approximately 3K optimization steps under the text-based reward system described earlier. All training is conducted on 8 H100 GPUs using the DeepSpeed ZeRO-3 optimization framework.

During evaluation, we apply our prompt rewriting method only to models that are either open-weight (Wan2.2) or allow disabling their built-in Prompt Extension module (Wan2.5 and Wan2.6-Flash). For the latter two models, we additionally report results with their native Prompt Extension (PE) enabled to provide a direct comparison. Beyond VWG-Bench, we further evaluate Vid-PRE on two additional high-quality open benchmarks. The first is MME-CoF, which contains 59 high-quality (Image, Prompt) pairs and evaluates generated videos from five aspects: instruction alignment, temporal consistency, visual stability, content fidelity, and focus relevance. The second benchmark is V-ReasonBench, which consists of 12 task categories with a total of 327 (Image, Prompt, Ground Truth) triplets. Each category is designed to be last-frame verifiable, meaning that task success can be determined solely from the final frame (e.g., Sudoku or arithmetic reasoning tasks). The benchmark groups the 12 categories into four reasoning dimensions: Structured Problem-Solving, Spatial Cognition, Pattern-based Inference, and Physical Dynamics. We report the mean score over five random seeds for each dimension.

Table 2: Baseline diagnostic on `VWG-Bench`. All models are evaluated without prompt extension. Commercial models are included to contextualize task difficulty.

Model	Video Quality	Prog. Consist.	Rule Follow.	Prog. Goal	Last Goal	Overall
Wan2.2	2.32	2.33	1.56	2.45	1.98	2.13
Wan2.5	2.70	2.86	3.14	2.48	2.56	2.75
Wan2.6-Flash	2.73	2.92	3.23	2.50	2.62	2.80
Kling-2.5-Turbo-Pro	2.95	3.12	3.38	2.62	2.78	2.97
Sora2	3.45	3.68	3.79	3.32	2.98	3.44
Veo3.1	**3.54**	**3.78**	**3.95**	**3.45**	**3.05**	**3.55**

4.2 Main Results

Benchmarking Current Models on `VWG-Bench`. We first establish a baseline diagnostic using VWG-Bench (Table 2). Even the strongest commercial model (Veo3.1) achieves only 3.55 overall on a 1–5 scale, underscoring the challenge of our benchmark. A consistent pattern emerges across all models: video quality scores are substantially higher than reasoning-intensive metrics. For example, Wan2.2 attains 2.32 in video quality but only 1.56 in implicit rule following, and even Sora2 drops from 3.45 to 2.98 between video quality and last-frame goal accuracy. This confirms our central hypothesis: current models excel at visual rendering but struggle to maintain physical constraints, logical rules, and temporal causality without explicit textual guidance.

Enhancing Reasoning with `Vid-PRE`. We evaluate Vid-PRE on two external benchmarks—MME-CoF and V-ReasonBench—as well as our VWG-Bench (Table 3). Integrating Vid-PRE yields substantial improvements across all models without any architectural modifications to the video generators. On MME-CoF (Table 3a), Vid-PRE boosts Wan2.2's overall score from 1.72 to 2.58, a 50% gain, with instruction alignment and content fidelity showing the largest absolute gains. It also outperforms Wan PE, raising Wan2.5 from 2.84 to 3.07 and Wan2.6-Flash from 2.82 to 3.09. On V-ReasonBench, the advantage is even more pronounced (Table 3b), whose tasks demand fine-grained physical and spatial reasoning. Vid-PRE improves Wan2.2's average accuracy from 26.65 to 44.48, a 67% gain, with spatial cognition increasing from 1.32 to 36.25. Against Wan PE, Vid-PRE delivers an additional +25% gain on Wan2.5 and +33% on Wan2.6-Flash, confirming that generic prompt rewriting is no substitute for structured, constraint-aware reasoning. On VWG-Bench (Table 3c), Vid-PRE raises Wan2.2 from 2.13 to 2.51, an 18% gain, and improves its weakest baseline dimension, implicit rule following, from 1.56 to 2.12. For Wan2.6-Flash, Vid-PRE achieves 3.26, surpassing the Wan PE baseline of 3.04 and narrowing the gap to commercial systems like Sora2 (3.46) and Veo3.1 (3.55).

4.3 Analysis and Discussion

Evaluator Reliability. We validate our VLM-as-Judge protocol by comparing automatic scores with human judgments. Specifically, we sample 100 prompts across all nine VWG-Bench dimensions and evaluate 200 videos generated by

Table 3: Performance of `Vid-PRE` across reasoning benchmarks. Wan PE denotes the model's built-in prompt extension. `Vid-PRE` consistently outperforms both the base models and their native PE across all metrics.

(a) MME-CoF.

Model	Prompt Extension	Instr. Align.	Temp. Consist.	Vis. Stability	Content Fidelity	Focus Relevance	Overall Score
Wan2.2	None	0.69	2.36	2.63	1.16	1.77	1.72
	`Vid-PRE`	**1.65**	**2.95**	**3.11**	**2.38**	**2.82**	**2.58**
Wan2.5	None	1.57	2.85	3.34	2.45	2.87	2.62
	Wan PE	1.98	2.88	3.41	2.59	3.35	2.84
	`Vid-PRE`	**2.34**	**3.28**	**3.59**	**2.69**	**3.47**	**3.07**
Wan2.6-Flash	None	1.67	3.05	3.39	2.49	2.85	2.69
	Wan PE	2.00	2.85	3.34	2.61	3.31	2.82
	`Vid-PRE`	**2.25**	**3.26**	**3.62**	**2.86**	**3.43**	**3.09**

(b) V-ReasonBench.

Model	Prompt Extension	Struct. Solving	Spatial Cognition	Pattern Inference	Physical Dynamics	Average
Wan2.2	None	29.27	1.32	45.12	30.90	26.65
	`Vid-PRE`	**33.90**	**36.25**	**60.49**	**47.29**	**44.48**
Wan2.5	None	25.95	20.05	46.30	33.05	31.34
	Wan PE	35.55	38.25	52.30	38.15	41.06
	`Vid-PRE`	**44.05**	**42.35**	**66.15**	**52.20**	**51.19**
Wan2.6-Flash	None	26.85	20.71	47.10	33.74	32.10
	Wan PE	36.57	29.18	53.23	39.02	39.50
	`Vid-PRE`	**45.46**	**43.78**	**67.35**	**53.68**	**52.57**

(c) VWG-Bench.

Model	Prompt Extension	Video Quality	Prog. Consist.	Rule Follow.	Prog. Goal	Last Goal	Overall
Wan2.2	None	2.32	2.33	1.56	2.45	1.98	2.13
	`Vid-PRE`	**2.65**	**2.68**	**2.12**	**2.75**	**2.35**	**2.51**
Wan2.5	None	2.70	2.86	3.14	2.48	2.56	2.75
	Wan PE	3.01	3.18	3.35	2.66	2.74	2.99
	`Vid-PRE`	**3.25**	**3.41**	**3.55**	**2.83**	**2.93**	**3.19**
Wan2.6-Flash	None	2.73	2.92	3.23	2.50	2.62	2.80
	Wan PE	3.05	3.18	3.45	2.68	2.86	3.04
	`Vid-PRE`	**3.28**	**3.49**	**3.62**	**2.88**	**3.02**	**3.26**

Veo3.1 and Wan2.6-Flash. Three human experts independently score each video on the same 1–5 criteria used by the automatic evaluator. As shown in Table 4, Spearman's ρ ranges from 0.69 to 0.80 and the mean absolute error remains below 0.55 across all criteria. This agreement indicates that the automatic scores track human-perceived reasoning quality rather than only superficial rendering preference, supporting the reliability of the benchmark-level model comparisons.

Ablation Study. We ablate the contribution of each training stage on both external benchmarks using Wan2.2 and Wan2.6-Flash as base generators (Table 5). SFT alone already provides a significant lift over the baselines, showing that structured reasoning and constraint-aware prompt generation are effective even without reinforcement learning. The subsequent GRPO stage consistently pushes performance further, as our text-based reward system explicitly penalizes spatial hallucinations and logical misalignments. The same trend holds for Wan2.6-Flash. These results confirm that the two stages are complementary:

Table 4: Agreement between human experts and Gemini-3.0-Pro on the VWG-Bench evaluation criteria.

Criterion	Human	Gemini	$\rho\uparrow$	MAE$\downarrow$
Video Quality	3.20	3.44	0.69	0.54
Progress Consistency	3.42	3.35	0.78	0.42
Rule Following	3.47	3.59	0.73	0.48
Progress Goal	3.06	2.98	0.75	0.41
Last-Frame Goal	2.93	2.84	0.80	0.39

SFT provides the structural foundation, and GRPO refines the policy toward tighter constraint satisfaction.

Table 5: Ablation on training stages.

Model	Variant	MME-CoF	V-ReasonBench
Wan2.2	Base	1.72	26.65
	+ Vid-PRE$_{\text{SFT}}$	2.44	40.26
	+ Vid-PRE$_{\text{SFT+RL}}$	**2.58**	**44.48**
Wan2.6-Flash	Base	2.69	32.10
	+ Vid-PRE$_{\text{SFT}}$	2.97	45.96
	+ Vid-PRE$_{\text{SFT+RL}}$	**3.09**	**52.57**

Qualitative Analysis. Figure 3 presents a qualitative comparison of video generation with and without Vid-PRE across diverse reasoning tasks. As illustrated, base video models fundamentally lack the capacity for zero-shot logical deduction. When given standard prompts, even those explicitly containing strict negative constraints, the base models suffer from severe reasoning failures and visual degradation. For instance, they hallucinate unrelated objects in visual analogies. By contrast, Vid-PRE dramatically enhances both reasoning capability and generation quality by transforming abstract logical problems into explicit, renderable visual instructions. It first employs a CoT to autonomously deduce the correct logical solutions and then rewrites the prompt to provide precise, step-by-step visual guidance. Consequently, Vid-PRE ensures strict physical constraint compliance, prevents visual hallucinations, and successfully executes complex think-with-video tasks.

5 Related Work

Image-to-Video Generation. I2V generation has progressed rapidly, driven primarily by diffusion and transformer architectures. Since Stable Video Diffusion [2] established the paradigm of image-conditioned generation, subsequent works have scaled this approach considerably. Open-source frameworks such as

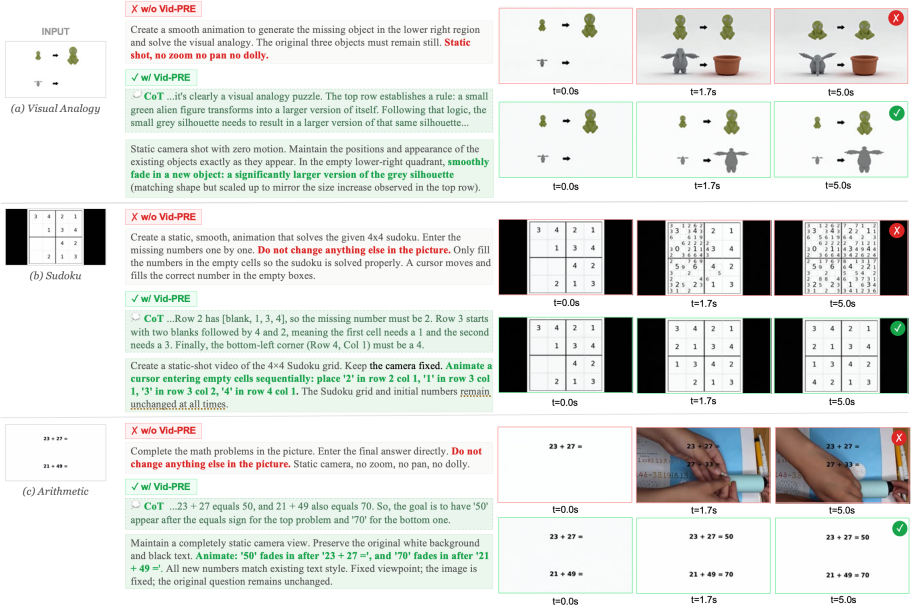

Fig. 3: Qualitative comparison of video generation with and without `Vid-PRE`. While generic constraints (red) fail to prevent logical errors and visual hallucinations in base models, `Vid-PRE` translates explicit reasoning into precise visual instructions (green), significantly improving both reasoning accuracy and video quality.

CogVideoX [41], Wan [35], and HunyuanVideo [20] have democratized high-resolution synthesis through advanced expert transformers and 3D causal VAEs, while commercial systems like Sora [31], Veo [14], Kling [21], and Seedance [12] have achieved photorealistic fidelity and remarkable temporal consistency. Despite these advances in rendering visual dynamics, their capacity for structured, rule-bound reasoning remains fundamentally underexplored.

Prompt Rewriting and Optimization. The distributional gap between concise user inputs and the detailed captions used during model training makes prompt rewriting critical for video generation. Early approaches bridge this gap via zero-shot in-context learning using LLMs (*e.g.*, GLM-4 in CogVideoX [41] or GPT-4o in Open-Sora [43]). To mitigate hallucination and intent drift inherent in zero-shot rewriting, recent methods have shifted toward reward-guided optimization, employing techniques such as Direct Preference Optimization (Prompt-A-Video [19]), multi-level feedback (VPO [5]), and retrieval-augmented refinement (RAPO [11]). However, these techniques predominantly focus on enhancing aesthetic quality and generic text-video alignment. In contrast, Vid-PRE specifically targets reasoning enhancement by offloading the cognitive burden to a VLM and training with purely text-based reinforcement learning signals.

Benchmarks for I2V Models. Existing evaluation frameworks broadly fall into two categories. The first focuses on perceptual quality and general gen-

eration capabilities: benchmarks like VBench/VBench++ [17, 18] and WorldScore [8] assess visual fidelity and scene dynamics but do not probe higher-order logical reasoning. The second category, motivated by recent demonstrations of reasoning in video generation [38], evaluates the cognitive limits of video models. Several broad benchmarks have emerged (*e.g.*, MME-CoF [15], Video-ThinkBench [34], TiViBench [4]), alongside specialized ones targeting procedural pathfinding (VR-Bench [40]), strict rule execution (RULER-Bench [16]), and implicit physical and visual rules (*e.g.*, V-ReasonBench [29], RISE-Video [25], MMGR [3]), as shown in Table 1. Despite advancing the evaluation landscape, these benchmarks often isolate specific cognitive skills or rely on holistic, coarse-grained scoring. VWG-Bench addresses these limitations with a comprehensive taxonomy spanning 9 dimensions and 38 tasks, equipped with a fine-grained three-level VLM-as-Judge framework.

6 Conclusion

In this work, we introduced `VWG-Bench`, a comprehensive evaluation framework designed to rigorously assess the think-with-video reasoning capabilities of modern video generative models. Through our three-level VLM-as-Judge evaluation, we revealed a persistent gap: current models excel at visual rendering but struggle with logic-heavy, rule-constrained, and long-horizon tasks. To bridge this gap, we proposed `Vid-PRE`, a model-agnostic prompt enhancer that offloads spatio-temporal reasoning to a dedicated VLM and translates raw instructions into constraint-aware, executable prompts. Trained via SFT and GRPO with purely text-based rewards, `Vid-PRE` yields substantial and consistent improvements in reasoning accuracy across multiple state-of-the-art generators.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
3. Cai, Z., Qiu, H., Ma, T., Zhao, H., Zhou, G., Huang, K.H., Kordjamshidi, P., Zhang, M., Xiao, W., Gu, J., et al.: Mmgr: Multi-modal generative reasoning. arXiv preprint arXiv:2512.14691 (2025)
4. Chen, H.H., Lan, D., Shu, W.J., Liu, Q., Wang, Z., Chen, S., Cheng, W., Chen, K., Zhang, H., Zhang, Z., et al.: Tivibench: Benchmarking think-in-video reasoning for video generative models. arXiv preprint arXiv:2511.13704 (2025)
5. Cheng, J., Lyu, R., Gu, X., Liu, X., Xu, J., Lu, Y., Teng, J., Yang, Z., Dong, Y., Tang, J., et al.: Vpo: Aligning text-to-video generation models with prompt optimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15636–15645 (2025)

6. Cho, J., Puspitasari, F.D., Zheng, S., Zheng, J., Lee, L.H., Kim, T.H., Hong, C.S., Zhang, C.: Sora as an agi world model? a complete survey on text-to-video generation. arXiv preprint arXiv:2403.05131 **1**(4) (2024)
7. Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al.: Understanding world or predicting future? a comprehensive survey of world models. ACM Computing Surveys **58**(3), 1–38 (2025)
8. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27713–27724 (2025)
9. Fang, X., Ma, L., Chen, Z., Zhou, M., Qi, G.j.: Inflvg: Reinforce inference-time consistent long video generation with grpo. arXiv preprint arXiv:2505.17574 (2025)
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
11. Gao, B., Gao, X., Wu, X., Zhou, Y., Qiao, Y., Niu, L., Chen, X., Wang, Y.: The devil is in the prompts: Retrieval-augmented prompt optimization for text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3173–3183 (2025)
12. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
13. Ghazanfari, S., Croce, F., Flammarion, N., Krishnamurthy, P., Khorrami, F., Garg, S.: Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. arXiv preprint arXiv:2506.00318 (2025)
14. Google DeepMind: Veo-3 technical report. Tech. rep., Google DeepMind (May 2025), `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`, accessed 25 June 2026
15. Guo, Z., Chen, X., Zhang, R., An, R., Qi, Y., Jiang, D., Li, X., Zhang, M., Li, H., Heng, P.A.: Are video models ready as zero-shot reasoners? an empirical study with the mme-cof benchmark. arXiv preprint arXiv:2510.26802 (2025)
16. He, X., Fan, Z., Li, H., Zhuo, F., Xu, H., Cheng, S., Weng, D., Liu, H., Ye, C., Wu, B.: Ruler-bench: Probing rule-based reasoning abilities of next-level video generation models for vision foundation intelligence. arXiv preprint arXiv:2512.02622 (2025)
17. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
18. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
19. Ji, Y., Zhang, J., Wu, J., Zhang, S., Chen, S., Ge, C., Sun, P., Chen, W., Shao, W., Xiao, X., et al.: Prompt-a-video: Prompt your video diffusion model via preference-aligned llm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18725–18735 (2025)
20. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
21. Kuaishou Technology: Kling ai: Next-generation ai creative studio. `https://klingai.com/` (June 2024), accessed 25 June 2026

22. LeCun, Y., et al.: A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review **62**(1), 1–62 (2022)
23. Li, M., Zhong, J., Zhao, S., Zhang, H., Lin, S., Lai, Y., Wei, C., Psounis, K., Zhang, K.: Tir-bench: A comprehensive benchmark for agentic thinking-with-images reasoning. arXiv preprint arXiv:2511.01833 (2025)
24. Lin, M., Wang, X., Wang, Y., Wang, S., Dai, F., Ding, P., Wang, C., Zuo, Z., Sang, N., Huang, S., et al.: Exploring the evolution of physics cognition in video generation: A survey. arXiv preprint arXiv:2503.21765 (2025)
25. Liu, M., Ma, S., Meng, S., Zhao, X., Zhang, Z., Zhang, S., Zhong, Z., Chen, P., Cao, H., Sun, X., et al.: Rise-video: Can video generators decode implicit world rules? arXiv preprint arXiv:2602.05986 (2026)
26. Liu, X., Xu, Z., Li, M., Wang, K., Lee, Y.J., Shang, Y.: Can world simulators reason? gen-vire: A generative visual reasoning benchmark. arXiv preprint arXiv:2511.13853 (2025)
27. Luo, M., Li, B., Xu, S., Zhang, S., Chen, Q., Han, M., Chen, W., Huang, Y., Fei, H., Lee, M.L., et al.: Unveiling the cognitive compass: Theory-of-mind-guided multimodal emotion reasoning. arXiv preprint arXiv:2602.00971 (2026)
28. Luo, M., Wu, S., Jing, L., Ju, T., Zheng, L., Lai, J., Wu, T., Du, X., Li, J., Yan, S., et al.: Dr. v: A hierarchical perception-temporal-cognition framework to diagnose video hallucination by fine-grained spatial-temporal grounding. International Journal of Computer Vision **134**(6), 278 (2026)
29. Luo, Y., Zhao, X., Lin, B., Zhu, L., Tang, L., Liu, Y., Chen, Y.C., Qian, S., Wang, X., You, Y.: V-reasonbench: Toward unified reasoning benchmark suite for video generation models. arXiv preprint arXiv:2511.16668 (2025)
30. Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M., Callison-Burch, C.: Faithful chain-of-thought reasoning. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 305–329 (2023)
31. OpenAI: Sora 2 system card. Tech. rep., OpenAI (September 2025), `https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf`, accessed 25 June 2026
32. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
33. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., et al.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918 (2025)
34. Tong, J., Mou, Y., Li, H., Li, M., Yang, Y., Zhang, M., Chen, Q., Liang, T., Hu, X., Zheng, Y., et al.: Thinking with video: Video generation as a promising multimodal reasoning paradigm. arXiv preprint arXiv:2511.04570 (2025)
35. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
36. Wang, Y., Liu, X., Pang, W., Ma, L., Yuan, S., Debevec, P., Yu, N.: Survey of video diffusion models: Foundations, implementations, and applications. arXiv preprint arXiv:2504.16081 (2025)
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

38. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
39. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
40. Yang, C., Wan, H., Peng, Y., Cheng, X., Yu, Z., Zhang, J., Yu, J., Yu, X., Zheng, X., Zhou, D., et al.: Reasoning via video: The first evaluation of video models' reasoning abilities through maze-solving tasks. arXiv preprint arXiv:2511.15065 (2025)
41. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
42. Zhang, X., Du, C., Pang, T., Liu, Q., Gao, W., Lin, M.: Chain of preference optimization: Improving chain-of-thought reasoning in llms. Advances in Neural Information Processing Systems **37**, 333–356 (2024)
43. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
44. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
45. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)