How Far Are Video Models from True Multimodal Reasoning?

Xiaotian Zhang[1,2] ♢ ⋆, Jianhui Wei[1,2] ♢, Yuan Wang[1] ♢, Jie Tan[1], Yichen Li[1], Yan Zhang[2] ♠, Ziyi Chen[2], Daoan Zhang[2], Dezhi Yu[2], Wei Xu[2], Songtao Jiang[1], and Zuozhu Liu[1] ♣

[1] Zhejiang University
[2] ByteDance
♢Equal Contribution. ♠Project Leader. ♣Corresponding Author.
{xiaotian.24, jianhui1.24}@intl.zju.edu.cn

Abstract. Despite remarkable progress toward general-purpose video models, a critical question remains unanswered: how far are these models from achieving true multimodal reasoning? Existing benchmarks fail to address this question rigorously, as they remain constrained by straightforward task designs and fragmented evaluation metrics that neglect complex multimodal reasoning. To bridge this gap, we introduce CLVG-Bench, an evaluation framework designed to probe video models' zero-shot reasoning capabilities via **C**ontext **L**earning in **V**ideo **G**eneration. CLVG-Bench comprises more than 1,000 high-quality, manually annotated metadata across 6 categories and 47 subcategories, covering complex scenarios including physical simulation, logical reasoning, and interactive contexts. To enable rigorous and scalable assessment, we further propose an **A**daptive **V**ideo **E**valuator (AVE) that aligns with human expert perception using minimal annotations, delivering interpretable textual feedback across diverse video context tasks. Extensive experiments reveal a striking answer to our central question: while state-of-the-art (SOTA) video models, such as Seedance 2.0, demonstrate competence on certain understanding and reasoning subtasks, they fall substantially short with logically grounded and interactive generation tasks (achieving success rates $< 25\%$ and $\sim 0\%$, respectively), exposing multimodal reasoning and physical grounding as critical bottlenecks. By systematically quantifying these limitations, the proposed method provides actionable feedbacks and a clear roadmap toward truly robust, general-purpose video models. CLVG-Bench and code are released at here.

Keywords: Video generation · Multimodal reasoning · Video evaluation

1 Introduction

Driven by large-scale training on web-scale data with generative objectives [6,27, 28, 64], video models have demonstrated groundbreaking zero-shot capabilities,

⋆ Work done during internship at ByteDance.

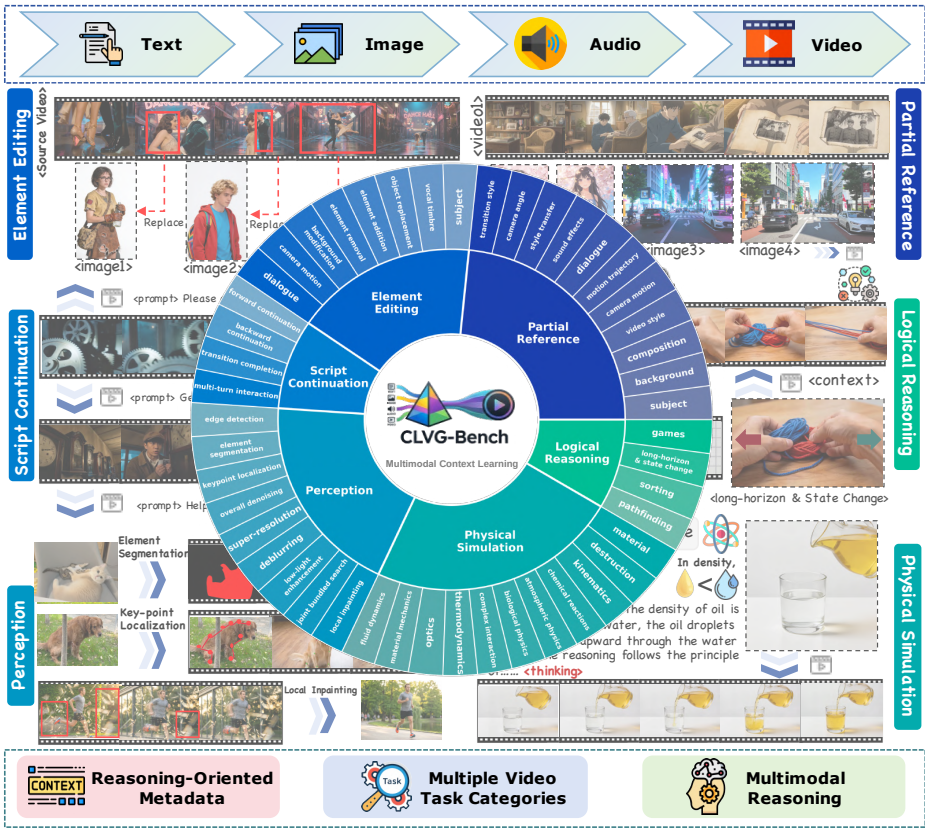

Fig. 1: Overview of CLVG-Bench. Our framework evaluates video models via Context Learning in Video Generation across arbitrary modalities (Text, Image, Audio, Video). It systematically probes reasoning capabilities across **6** categories and **47** subcategories, which provides a roadmap for general-purpose video models.

evolving from mere instruction following to complex understanding and reasoning [49,53,62,66,78]. Specifically, traditional reference-based video generation has transitioned from simple, single-reference conditioning [29,32,33,35,45,56,74] to accommodating multi-subject and multimodal references [22,63]. Concurrently, the primary focus of these tasks is shifting from prioritizing sheer visual fidelity to emphasizing broader reasoning capabilities. Despite this immense potential and the surging interest in video reasoning, the community still lacks a systematic formulation and comprehensive benchmarking for these emerging paradigms. To address this, we formalize the task of reasoning-driven video generation, conditioned on arbitrary combinations of modalities (text, image, audio, and video), as **Context Learning in Video Generation**.

Second, video quality evaluation has transitioned from fragmented scoring metrics [24] toward fine-grained and unified textual interpretations. Previous

evaluation paradigms typically rely on task-specific metric combinations [9, 23, 26, 39, 40, 50, 67, 79] or reward models trained on multi-dimensional preference datasets [3,4,19,20,70]. However, fragmented and coarse metrics often fail to provide actionable textual feedback, while reward model training is frequently bottlenecked by the scarcity of high-quality data and remains susceptible to reward hacking [10]. To circumvent these issues, recent agent-based frameworks [17, 65, 72] have been adapted for unified video evaluation, offering both quantitative scores and qualitative feedback. Nonetheless, their reliance on hand-crafted system prompts and static in-context examples severely constrains their generalization and reproducibility. They primarily focus on visual fidelity while overlooking intrinsic logical reasoning within videos. Consequently, efficiently aligning these agents to new video evaluation tasks with minimal expert supervision remains a critical yet underexplored problem.

We introduce **CLVG-Bench**, designed to comprehensively evaluate video models' ability to perform zero-shot reasoning on multimodal contexts. We systematically define 6 fundamental categories and 47 subcategories, and have carefully curated more than 1,000 high-quality metadata entries with video reasoning-oriented contexts, including text instruction, multi-shot and multi-subject videos, reference images and audio. To prevent the copyright issues and data contamination in evaluation, all metadata (*e.g.* videos, images) are manually curated. We further propose an **A**daptive **V**ideo **E**valuator (**AVE**), leveraging an Automatic Prompt Optimization (APO) framework equipped with a novel semantic matching function for open-ended text evaluation. This mechanism empowers the evaluation agent to adapt to specific tasks using minimal expert annotations while simultaneously generating detailed textual feedback. Consequently, AVE serves as a interpretable metric for assessing the reasoning capabilities of video models, supporting scalable evaluation across diverse models and tasks, and facilitating fine-grained error attribution.

Experimental results reveal significant performance variations among current video models across different reasoning tasks. In particular, these models exhibit considerable shortcomings in tasks related to physical reasoning, logical inference, and interactive generation. This highlights a critical gap in their capability to simulate and reason about real-world dynamics.

Further analysis reveals that explicit context understanding provided by vision-language models aids video models in generating physically plausible content. This indicates that current video models lack the intrinsic ability to autonomously infer physical causality within instructions. Moving forward, future architectures should integrate comprehension and generation components more seamlessly to achieve robust, reliable reasoning capabilities in complex scenarios.

Our major contributions are threefold:

- We introduce CLVG-Bench, a novel evaluation framework that abstracts video generation tasks into the definition of context learning video generation, and systematically evaluates the capability of current video models in simulating and reasoning about real-world dynamics.

- We propose the Adaptive Video Evaluator, a flexible evaluation framework designed for open-ended generation tasks. This evaluator dynamically adjusts based on a minimal set of human annotations, offering a versatile approach to assess tasks with varying contexts.
- Our study uncovers the limitations of current video models in multimodal reasoning. We advocate for a tighter integration of understanding and generation to enhance model performance in complex video generation scenarios.

2 Related Works

2.1 Evolution of Video Models

Video models have evolved from early diffusion frameworks [56, 57, 69] to advanced Diffusion Transformer (DiT) architectures [12,34,44,53,55,80], achieving superior cinematic quality, physical simulation, and efficiency [16]. To bridge understanding and generation, unified models [52,61] now integrate cross-modal decoding within single architectures. Recent works advance this bidirectional paradigm through joint audio-visual DiTs [48] and integrating LLMs with 3D visual tokenizers [8,37,63], driving holistic multimodal intelligence.

2.2 Video Benchmark

Video benchmarks have shifted from discriminative understanding to generative reasoning [14]. Early efforts [7, 31, 71] focused on temporal and diagnostic reasoning, a trajectory recently extended by VBVR [58] to probe complex logical chains. However, these web-scraped datasets face copyright and contamination risks while prioritizing perception over synthesis. Consequently, generation and editing benchmarks [18,25,46,51] established aesthetic and instruction-following frameworks. Although recent works [29, 75] support reference-guided editing, they are often limited to single-shot scenarios and lack cross-shot consistency. More recently, the field has gravitated toward physical and spatial grounding. Benchmarks such as PhyGenBench [41], RBench-V [13], and SpatialVizBench [60] evaluate physical commonsense and 3D spatial relationships. Despite these advances, existing frameworks typically evaluate tasks in isolation, lacking a unified paradigm for multimodal context video generation.

2.3 Video Evaluation Methods

Existing video generation evaluation methodologies can be broadly categorized into three paradigms: heuristic foundation metrics, trained reward models, and agent-based LLM-as-judge frameworks. Heuristic metrics based on vision foundation models [23, 26, 39, 50] are widely adopted for their computational efficiency. For example, CLIP-Score and its variants [21, 40] for text-to-video consistency measurement. While efficient, these methods only yield coarse holistic scores and lack the fine-grained feedback required for detailed error analysis.

Table 1: Comparison of CLVG-Bench with existing video benchmarks. Our benchmark support arbitrary combinations of multimodal inputs (Text, Image, Audio, and Video) while covering the broadest range of reasoning-oriented generation tasks, from element editing to multi-turn interaction.

Tasks : ①: E.E.	②: P.R.	③: S.C.	④: P.S.	⑤: Perc.	⑥: L.R.	⑦: Interaction	
Benchmark	**Text**	**Image**	**Audio**	**Video**	**Reasoning**	**Count**	**Categories**
VBench [25]	✓	✗	✗	✗	✗	800	②
VE-Bench [51]	✓	✗	✗	✓	✗	169	① ②
UNIC [76]	✓	✓	✗	✓	✗	210	① ②
VACE-Bench [30]	✓	✓	✗	✓	✗	240	① ②
FiVE-Bench [36]	✓	✗	✗	✓	✗	420	① ②
UniVBench [65]	✓	✓	✗	✓	✗	620	① ②
RBench-V [13]	✓	✓	✗	✗	✓	803	⑥
Video-CraftBench [47]	✓	✓	✗	✓	✓	150	⑥
OpenVE-3M [18]	✓	✗	✗	✓	✗	431	① ②
VBVR-Bench [59]	✓	✓	✗	✓	✓	7500	④ ⑤ ⑥
CLVG-Bench	✓	✓	✓	✓	✓	1390	①∼⑦

Note: E.E.: Element Editing; **P.R.**: Partial Reference; **S.C.**: Script Continuation; **P.S.**: Physical Simulation; **Perc.**: Perception; **L.R.**: Logical Reasoning.

The second paradigm focuses on trained specialized reward models. For example: [20] which uses Video-LLMs to generate human-aligned multi-dimensional regression scores. However, these models require substantial expert-annotated data, making them hard to scale with rapidly evolving video tasks. Recently, agent-based LLM-as-judge frameworks have emerged as flexible evaluation solutions [17, 65, 72]. Despite their versatility, these agent-based approaches rely on expert-crafted system prompts and in-context examples, limiting reproducibility and scalability. To address these inherent limitations, our AVE leverages prompt optimization to iteratively refine the evaluation agent's system prompt, enabling dynamic adaptation to diverse video tasks with minimal expert annotations.

3 Method

3.1 Dataset Construction

Metadata Definition. Irrespective of the specific downstream tasks, user inputs can be systematically categorized based on their modality formats. As delineated in Tab. 1, in a distinct departure from prior works, CLVG-Bench abstracts user prompts into versatile combinations of four foundational modalities: text, image, audio, and video. Furthermore, adhering to a reasoning-centric paradigm, we incorporate highly intricate reference-guided editing scenarios, including multi-shot and multi-subject tasks. Based on partial definitions from prior works [47, 63, 66], we categorize context in video generation tasks into six fundamental categories comprising 47 fine-grained subcategories. The traditional elements editing, partial reference, and script continuation are seamlessly integrated with more context-aware physical simulation, perception, logical reasoning, and multi-turn interaction to formulate the comprehensive metadata architecture of CLVG-Bench (we temporarily incorporate the multi-turn interactive

generation into the script continuation. Fig. 1 illustrates the detailed categories and provides several example demonstrations).

Metadata Synthesis. Directly using copyrighted web data may not only contaminate the evaluation data but also introduce a series of potential privacy concerns. To address these issues, all content in CLVG-Bench is human-created and carefully curated, ensuring that it is entirely free of copyright restrictions. This design enables a fair evaluation of in-context video generation capabilities and instruction following performance without legal constraints.

As described above, the metadata consists of four components: text (also called context), images, video, and audio. The context serves as the fundamental basis for reference to video generation. We use Seed 2.0 Pro [48] to generate reference contexts covering all subcategories under the following principles: (1) extract themes for all six categories and their subcategories; (2) expand the original theme and further enrich it into realistic user-oriented requirement descriptions. Subsequently, three annotators independently verify the generated reference contexts according to their assigned categories, examining semantic completeness and requirement accuracy. Gemini 3 Pro [11] provides additional automated verification by cross-checking factual statements. Any context that fails to adhere to the predefined category specifications undergoes collaborative review, where annotators discuss discrepancies and produce a corrected version.

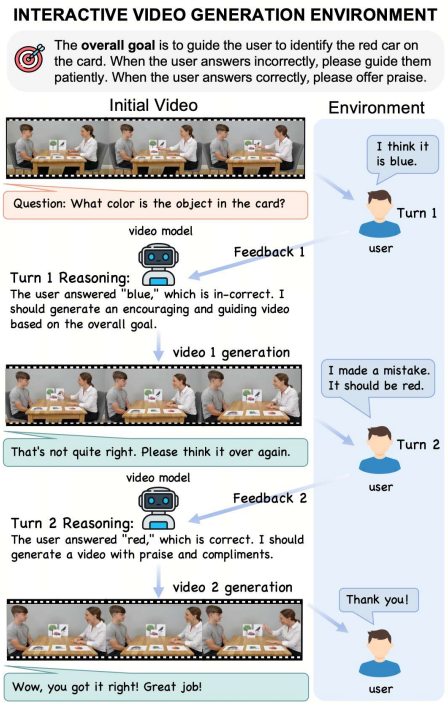

Fig. 2: An environment for testing the multi-turn interactive generation capabilities of video models, where the model is expected to infer and predict a suitable video script based on the history context and user feedback.

Before generating the metadata videos, we follow three steps to produce content-rich and diverse video scripts: (1) Based on Andrew *et al.* [2] and Bordwell *et al.* [5], we define five fundamental elements for video generation: subject, vocal characteristics of the subject, number of shots, camera movement, and video genre. (2) For each video script, we randomly assign five labels (sampled without replacement to ensure a uniform distribution) to construct its core

elements. (3) We then invoke Seed 2.0 Pro to generate a complete and coherent script for each case based on the assigned labels, ensuring that it is directly usable for video generation. Finally, we use Seedance 2.0 to generate the video conditioned on the full script.

UniVBench [65] contains a diverse set of reference images that are free from copyright concerns. We manually curated and filtered out samples that might cause style drift or introduce ambiguous references. The remaining images were retained and used as metadata images. For the metadata audio, we curate and collect audio data from the Seed-VC open-source repository [38], filtering out segments with noisy backgrounds and low quality.

Multi-turn Interaction. While some prior work has explored the physical simulation and logical reasoning capabilities of video models [42, 47, 59, 68], few studies focus on the multi-turn interactive generation of videos. This task requires video models to not only understand but also iteratively predict video scripts based on user feedback, rather than simply following current input instructions. In CLVG-Bench, we introduce the novel task of multi-turn interactive video generation and design a series of interactive feedback environments. As illustrated in Fig. 2, inspired by early education scenarios, we require video models to assume the role of a teacher in the overall goal. The environment provides random feedback based on its content. The baseline model is expected to reason and generate the correct video script in response to the feedback turn by turn.

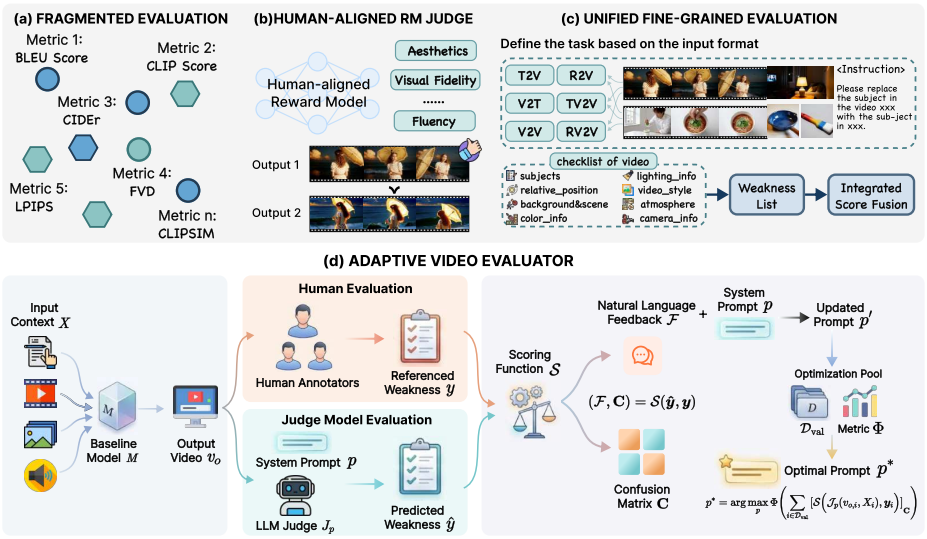

Fig. 3: The top tier represents the previous evaluation methodologies: (a) fragmented evaluation, (b) human-aligned reward models, and (c) unified fine-grained evaluation. The lower tier illustrates the pipeline of the Adaptive Video Evaluator.

3.2 Evaluation Pipeline

Problem Formulation. Let the input context be defined as a tuple $X \in \{\text{text}, v_r, \text{image}, \text{audio}\}$, where v_r represents a reference video. A baseline model $\mathcal{M}$ generates an output video v_o based on the input, denoted as $v_o = \mathcal{M}(X)$.

Human annotators assess v_o based on context X and provide a referenced set of noticeable weaknesses, denoted as $\boldsymbol{y}$. Concurrently, a VLM-based judge agent $\mathcal{J}_p$, conditioned on a system prompt p, predicts a corresponding set of weaknesses $\hat{\boldsymbol{y}} = \mathcal{J}_p(v_o, X)$. To evaluate the judge agent, a semantic matching function $\mathcal{S}$ compares the prediction $\hat{\boldsymbol{y}}$ against the referenced $\boldsymbol{y}$, yielding natural language feedback $\mathcal{F}$ and a normalized instance-level confusion matrix $\mathbf{C} \in [0,1]^{1\times 4}$ (comprising TP, TN, *etc.*):

$$(\mathcal{F}, \mathbf{C}) = \mathcal{S}(\hat{\boldsymbol{y}}, \boldsymbol{y}). \tag{1}$$

The feedback $\mathcal{F}$ details the discrepancies between $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$, while the elements of $\mathbf{C}$ sum to 1. The matching logic is shown in Algorithm 1.

To optimize the judge agent $\mathcal{J}_p$, we collect a batch of annotated dataset $\mathcal{D} = \{(X_i, v_{o,i}, \boldsymbol{y}_i)\}_{i=1}^{N}$. This dataset is further partitioned into training, validation, and testing subsets, denoted as $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, respectively.

During the optimization phase, we utilize $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ to explore and verify prompt candidates. By aggregating the instance-level confusion matrices across the validation set into a global confusion matrix, our objective is to find the optimal prompt p^* that maximizes a specific evaluation metric $\Phi(\cdot)$ (e.g., F1 score or Recall minus FPR):

$$p^* = \arg\max_{p} \Phi\left(\sum_{i \in \mathcal{D}_{\text{val}}} \Big[\mathcal{S}\Big(\mathcal{J}_p(v_{o,i}, X_i), \boldsymbol{y}_i \Big) \Big]_{\mathbf{C}} \right). \tag{2}$$

Finally, the optimal prompt p^* is frozen and evaluated on the unseen test set $\mathcal{D}_{\text{test}}$ to measure the generalized performance of the judge agent:

$$\text{Score}_{\text{test}} = \Phi\left(\sum_{i \in \mathcal{D}_{\text{test}}} \Big[\mathcal{S}\Big(\mathcal{J}_{p^*}(v_{o,i}, X_i), \boldsymbol{y}_i \Big) \Big]_{\mathbf{C}} \right). \tag{3}$$

Adaptive Video Evaluator. We introduce AVE, a prompt optimization framework designed to dynamically adapt to diverse video evaluation contexts. At the core of AVE lies a novel semantic matching function specifically engineered for open-ended, free-form evaluation. Departing from traditional APO techniques that rely on deterministic exact-matching, typical in QA or mathematical tasks, our approach conceptualizes evaluation as a semantic set-matching problem. This enables the model to align free-form human feedback with generated outputs, effectively disentangling complex textual comparisons to provide both robust quantitative metrics and actionable natural language insights.

Given a cost budget B, the AVE framework (Algorithm 2, Fig. 3: lower tier) iteratively optimizes the judge's prompt. In the beginning of each iteration, a

Algorithm 1 Semantic Matching $\mathcal{S}$

Require: referenced and predicted weaknesses $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$; semantic matching agent $Match$

function SEMANTICMATCHING($\boldsymbol{y}, \hat{\boldsymbol{y}}$)
  if $|\boldsymbol{y}| > 0 \wedge |\hat{\boldsymbol{y}}| > 0$ **then**
    $TP_{set} \leftarrow \boldsymbol{y} \cap_{Match} \hat{\boldsymbol{y}}$ ▷ Semantic set operations via $Match$
    $FN_{set} \leftarrow \boldsymbol{y} \setminus_{Match} TP_{set}$
    $FP_{set} \leftarrow \hat{\boldsymbol{y}} \setminus_{Match} TP_{set}$
    return $[|TP_{set}|, |FP_{set}|, |FN_{set}|]$
  end if
  return $[0, 0, 0]$
end function
$[n_{TP}, n_{FP}, n_{FN}] \leftarrow$ SEMANTICMATCHING($\boldsymbol{y}, \hat{\boldsymbol{y}}$)
$TP \leftarrow \mathbb{I}(|\boldsymbol{y}| > 0 \wedge |\hat{\boldsymbol{y}}| > 0) + n_{TP}$; $TN \leftarrow \mathbb{I}(|\boldsymbol{y}| = 0 \wedge |\hat{\boldsymbol{y}}| = 0)$
$FP \leftarrow \mathbb{I}(|\boldsymbol{y}| = 0 \wedge |\hat{\boldsymbol{y}}| > 0) + n_{FP}$; $FN \leftarrow \mathbb{I}(|\boldsymbol{y}| > 0 \wedge |\hat{\boldsymbol{y}}| = 0) + n_{FN}$
$N_{\text{total}} \leftarrow TP + TN + FP + FN$ ▷ Normalization
score $\leftarrow \left[\frac{TP}{N_{\text{total}}}, \frac{TN}{N_{\text{total}}}, \frac{FP}{N_{\text{total}}}, \frac{FN}{N_{\text{total}}}\right]$
if $FP = 0 \wedge FN = 0$ **then**
  feedback $\leftarrow$ "Perfect prediction."
else
  feedback $\leftarrow$ "Omissions: " $\oplus$ FN $\oplus$ "; Hallucinations: " $\oplus$ FP
end if
return feedback, score ▷ Corresponds to $\mathcal{F}_i$ and $\mathbf{C}_i$ in Algorithm 1

candidate prompt is sampled from the candidates (*e.g.* pareto sampling [1]). For mini-batch of data, $\mathcal{S}$ generates both a quantitative score for performance tracking and natural language feedback to guide the optimizer $\mathcal{O}$ in prompt refinement. This process ultimately returns the optimal prompt p^* that maximizes the global metric Φ on the validation set.

4 Experiments

4.1 Implementation Details

Baselines. For open-source video models, we use Wan2.2-14B [56] Hunyuan-Video1.5 [54], UniVideo [63], CogVideoX1.5-5B [73] and LTX-2 [15] following their official implementations and settings. For proprietary models, we use Seedance 2.0, Seedance 1.5 [49], Veo 3.1 Fast [12], Wan 2.7 and Sora2 [43].

Adaptive Video Evaluator. During the prompt optimization phase, we employ Seed 2.0 Pro [48] as the prompt optimizer and Seed 2.0 Lite as the judge model. We equip our semantic matching function to two popular APO techniques: TextGrad [77] and GEPA [1]. We set the total optimization budget to 30 US dollars, temperature to 0 and the maximum number of tokens to 32,000. To guarantee stable results, the judge model evaluates each instance five times, determining the final outcome via majority vote. The dataset consists of 600 annotated samples in total, evenly split into training, validation, and test sets.

Algorithm 2 AVE: Adaptive Video Evaluator

Require: Judge $\mathcal{J}_p$ with initial prompt p_0, datasets $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}$, semantic matching function $\mathcal{S}$, batch size b, budget B, prompt-score table $\mathcal{T}$, optimizer $\mathcal{O}$ (fixed meta-prompt), metric Φ

function EvaluateAndUpdate($p, \mathcal{D}$)
  for each $(X_i, v_{o,i}, \boldsymbol{y}_i) \in \mathcal{D}$ **do**
    $(\mathcal{F}_i, \mathbf{C}_i) \leftarrow \mathcal{S}\Big(\mathcal{J}_p(v_{o,i}, X_i), \boldsymbol{y}_i\Big)$
  end for
  $\text{score}_{\mathcal{D}} \leftarrow \Phi\big(\sum_{i\in\mathcal{D}} \mathbf{C}_i\big)$; $\text{feedback}_{\mathcal{D}} \leftarrow \sum_{i\in\mathcal{D}} \mathcal{F}_i$
  if $\mathcal{D} = \mathcal{D}_{\text{val}}$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{\langle \text{score}_{\mathcal{D}}, p\rangle\}$
  end if
  return $\text{score}_{\mathcal{D}}, \text{feedback}_{\mathcal{D}}$
end function
EvaluateAndUpdate($p_0, \mathcal{D}_{\text{val}}$) ▷ Initialize table $\mathcal{T}$ with p_0
while budget B not exhausted **do**
  $p \leftarrow$ SelectPrompt($\mathcal{T}$) ▷ Sample best candidate or use pareto sampling [1]
  $\mathcal{D}_{\text{mini}} \sim \mathcal{D}_{\text{train}}$ ▷ Sample minibatch of size b
  score, feedback $\leftarrow$ EvaluateAndUpdate($p, \mathcal{D}_{\text{mini}}$)
  $p' \leftarrow \mathcal{O}(p, \text{feedback})$
  EvaluateAndUpdate($p', \mathcal{D}_{\text{val}}$)
end while
return p^* with maximum score on $\mathcal{D}_{\text{val}}$ from $\mathcal{T}$

Finally, we adopt Recall minus False Positive Rate (Rec-FPR.), F1-Score (F1), Matthews Correlation Coefficient (MCC) as evaluation metrics. We reorganize the six categories into three broader training tasks: perception, prompt following (which includes element editing, partial reference, and script continuation), and physical and logical reasoning (covering both physical simulation and logical reasoning).

4.2 Experimental Results

Main Results. CLVG-Bench comprises a diverse spectrum of test cases with progressively increasing difficulty, ranging from standard text-to-video generation to more challenging settings involving multi-shot, multi-subject, and multi-video to reference. Considering that certain baseline models impose constraints on input formats or the number of references, we adopt a group-wise comparison protocol to ensure fairness; models within the same group are evaluated on an identical set of test cases. The quantitative results are reported in Tab. 2.

Empirical evaluations reveal that existing models consistently excel at conventional editing tasks (peaking at a 61.25% success rate), yet struggle significantly with tasks requiring more context-aware understanding and reasoning. Notably, even SOTA methods plateau at a marginal success rate of 21.25% on logical reasoning tasks. Furthermore, proprietary models deliver a marked performance margin over open-source video baselines on average. This discrepancy

Table 2: Main results on CLVG-Bench. We evaluate all baselines across six fundamental dimensions of CLVG-Bench. The evaluation is conducted by human annotators. Results are reported as pass rates (%).

Method	E.E.	P.R.	S.C.	P.S.	Perc.	L.R.
Proprietary Models						
Seedance 2.0* [48]	61.25	52.64	67.65	63.54	43.37	21.25
Sora 2† [43]	46.38	49.45	52.49	48.91	45.86	38.96
Seedance 1.5‡ [49]	-	35.47	-	37.50	40.00	8.75
Veo 3.1 Fast‡ [12]	-	46.38	-	46.88	36.00	16.25
Open Source Models						
UniVideo† [63]	8.1	8.06	4.08	10.00	3.00	3.75
LTX-2† [16]	0.72	11.29	2.04	16.00	2.00	2.50
Wan 2.2 14B‡ [56]	-	46.67	-	39.00	16.00	11.25
HunyuanVideo 13B‡ [33]	-	26.67	-	26.00	14.00	5.00

Note: *Supports video, audio, and image inputs; †Supports video and image inputs; ‡Supports only image inputs.

highlights the robust generalization capabilities of proprietary systems, rendering them inherently superior in scenarios with intricate reasoning demands. Concurrently, however, our findings expose lingering input constraints across current architectures: they frequently fall short of accommodating diverse reference combinations, thereby leaving a tangible gap between SOTA capabilities and the multifaceted demands of real-world users.

Prompt Optimization Results. As shown in Tab. 5, mainstream APO methods (TextGrad and GEPA) significantly enhance the judge model's evaluation performance compared to the vanilla baseline. For instance, TextGrad alone improves the MCC by up to 23.8% on the Perception task. Building upon these optimized prompts, our proposed SemanticMatch module consistently delivers further performance gains. On average, SemanticMatch provides an additional 4.9%, 1.7%, and 4.5% boost in MCC, F1 score, and Rec-FPR, respectively, across all tasks and baselines. Notably, it achieves a substantial improvement of 9.2% on the P.S.+L.R. task when integrated with TextGrad, and an 8.5% lift on the E.E.+P.R.+S.C. task for GEPA. Ultimately, the combined approach reaches a peak performance of 70.7% MCC on the Perception task, demonstrating the superior effectiveness of our semantic matching function.

Correlation with Human Judgement. To assess the correlation with human judgement , we report Kendall's τ in Tab. 3. Our prompt optimization leads to substantial performance gains across all task categories. Notably, the Seed 2.0 Lite model reaches a strong correlation of 0.707 in Perception tasks, while Seed

2.0 Pro achieves a high score of 0.620 in reasoning and simulation tasks after adaptation. These results demonstrate that optimization effectively enhances evaluator reliability, achieving high alignment with human preferences.

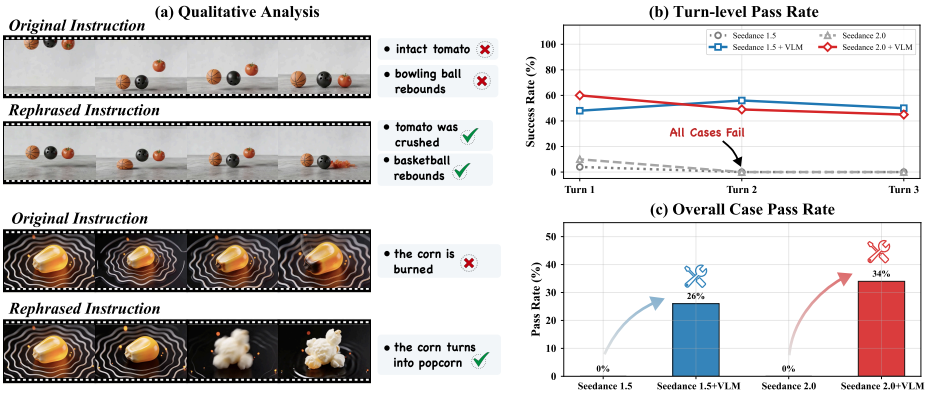

Fig. 4: (a) Qualitative comparison of videos generated from the original instructions and their rephrased counterparts. (b) Per-turn success rate in the multi-turn interaction setting. (c) Overall case pass rate (%).

4.3 Discussion and Analysis

Necessity of Context Learning. The results in Tab. 2 reveal that current video models still exhibit notable deficiencies in context learning and reasoning. A potential workaround is to leverage external understanding modules, such as VLMs, to assist understanding and video generation. In our experiments, we employ Seed 2.0 Pro to first interpret the provided context and then rewrite the initial instruction accordingly.

The quantitative and qualitative results are presented in Tab. 4 and Fig. 4 (a). With this auxiliary understanding step, the relatively weaker Seedance 1.5 shows a substantial improvement (+14.9% on Physical Simulation task) and even surpass Seedance 2.0 and Veo 3.1 Fast on Logical Reasoning task. However, this gain also suggests that external assistance only provides a partial remedy. Ultimately, enhancing the intrinsic reasoning capability of video models during training remains essential for addressing this limitation.

Struggle in Multi-turn Interaction. The multi-turn interactive generation task requires models to reason based on user feedback and implicit context, directly generating video content, which contrasts sharply with previous task formats. As shown in Fig. 4 (b), under the vanilla setting, top video models achieve a first-turn success rate of less than 10%, and completely fail during the second turn of interaction. This highlights their limitations in processing complex information and generating content based on reasoning.

Table 3: Evaluator performance on Kendall's τ on different models and tasks.

Model	E.E.+P.R. +S.C.	P.S. +L.R.	Perc.
Seed 2.0 Lite	0.239	0.368	0.433
+ Optimization	0.480 **+0.241**	0.577 **+0.209**	0.707 **+0.274**
Seed 2.0 Pro	0.00	0.462	0.379
+ Adaptation	0.273 **+0.273**	0.620 **+0.158**	0.440 **+0.061**

Table 4: Results (%) of **P.E.** and **L.R.** on VLM-enhanced script.

Model	P.S.	L.R.	AVG
Seedance 1.5	37.5	8.8	23.2
+ Seed 2.0 Pro	52.4 **+14.9**	23.8 **+15.0**	38.1 **+14.9**
Seedance 2.0	63.5	21.3	42.4
+ Seed 2.0 Pro	70.3 **+6.8**	49.0 **+27.7**	59.7 **+17.3**

Table 5: Evaluator performance on three types of tasks, results are presented as percentages. Adaptation means adapting optimized prompt to other models.

Model	E.E.+P.R.+S.C.			P.S.+L.R.			Perc.		
	MCC	F1	Rec-FPR	MCC	F1	Rec-FPR	MCC	F1	Rec-FPR
Seed 2.0 Lite	23.9	70.2	10.5	36.8	65.6	36.4	43.3	79.0	40.3
+ TextGrad	39.4	74.5	31.8	48.5	74.6	48.5	67.1	87.2	65.0
+ SemanticMatch	43.8	75.5	31.6	57.7	79.4	57.6	70.7	88.6	67.7
	+4.4	**+1.0**	**−0.2**	**+9.2**	**+4.8**	**+9.1**	**+3.6**	**+1.4**	**+2.7**
+ GEPA	39.3	74.1	26.3	54.6	76.9	54.6	64.3	86.4	60.0
+ SemanticMatch	47.8	76.9	36.8	51.6	75.0	51.5	70.7	88.6	67.7
	+8.5	**+2.8**	**+10.5**	**−3.0**	**−1.9**	**−3.1**	**+6.4**	**+2.2**	**+7.7**
Seed 2.0 Pro	0.00	67.8	0.00	46.2	75.9	42.4	37.9	78.7	29.2
+ Adaptation	17.8	69.1	10.8	62.0	82.2	60.6	44.0	80.0	38.1
	+17.8	**+1.3**	**+10.8**	**+15.8**	**+6.3**	**+18.2**	**+6.1**	**+1.3**	**+8.9**

However, when enhanced with the visual reasoning capabilities of VLM (Seed 2.0) to assist with context organization and script prediction, the first-turn success rates improve by 44% and 50%, respectively. Furthermore, the final pass rate after multiple rounds of interaction reaches 26% and 34%, as shown in Fig. 4 (c). Although these pass rates are still modest, they demonstrate that a stronger coupling of understanding and generation can significantly improve the generalization of video models to more realistic and dynamic scenarios.

Prompt Generalization. We found that the evaluation system prompt optimized with AVE can be directly transferred to other models and obtain performance gains. As shown in Tab. 5, when adapting the prompt optimized on Seed 2.0 Lite to the more capable Seed 2.0 Pro model, we observe consistent performance improvements across all three task families, with an average of 13.2% MCC, 3.0% F1, and 12.7% Rec-FPR gain on the three tasks. This shows that the optimized system prompt contains generalizable task-solving experience that can be shared across different models, enabling zero-cost cross-model adaptation with no additional optimization required.

Case Study on AVE. Fig. 5 illustrates a case study of prompt refinement utilizing our proposed AVE. Drawing inspiration from the human cognitive paradigm of learning from past failures, the process initiates with a straightforward, naïve instruction. By systematically diagnosing previous weaknesses, we iteratively

Physical Evaluation Initial Prompt (F1 Score: 65.6)

Your task is to evaluate whether the video contains any physical abnormality.

AVE Optimized Physical Evaluation Prompt (F1 Score: 79.4)

Your task is to evaluate whether the video contains any physical abnormality.
Follow these rules during evaluation:

1. **For any suspected physical abnormality**, cross-verify the state of relevant objects across multiple consecutive frames to confirm the abnormality is consistent, rather than a transient observation error.
2. **When evaluating mechanical interaction processes** (such as zipper operation, object assembly, ...
3. **Distinguish between real physical abnormalities and normal visual effects** caused by common .....
4. **When the video involves specific materials** (especially if the input prompt explicitly states the material type), verify that the material exhibits its corresponding unique physical characteristics .....
5. **Distinguish between actual physical changes of object attributes** (such as length, size, shape, etc.) and apparent visual changes caused by 3D perspective projection when objects move along the depth direction (towards or away from the camera) or rotate/change position relative to the camera. Before judging any observed attribute change as an abnormality, you must strictly follow these steps:
   - **a)** First identify at least two fixed non-moving reference objects in the scene (e.g. fixed supports, ...
   - **b)** Compare the target object's dimensions and position relative to these fixed reference objects across .....
   - **c)** Only when multiple fixed references confirm the change is consistent and .....
6. **Before starting evaluation, first extract the core action/process** described in the input prompt, ...
7. **When evaluating processes involving consumable objects** (e.g., chalk, pencils, erasers, candles, ink) .....
8. **When evaluating processes that produce uniform physical effects** on multiple relevant objects (e.g., ...
9. **For objects subjected to processes that can produce chemical or physical property changes** (e.g., ...
10. **After extracting the core action/process from the input prompt**, explicitly list all explicit action .....

Fig. 5: Case study of AVE-driven prompt refinement in the physical simulation task. This example illustrates the process of enhancing a basic initial prompt using AVE.

synthesize an empirical rubric. This refinement strategy yields a substantial performance gain (+13.8 in F1 score), firmly demonstrating that AVE exhibits robust generalizability to substantially more complex domains imposing stringent demands on contextual reasoning. Rather than merely fitting surface-level instructions, this iterative reflection empowers the model to deduce implicit relationships and maintain logical coherence across intricate, context-heavy tasks.

5 Conclusion

In this paper, we introduce CLVG-Bench, a benchmark designed to systematically evaluate multimodal reasoning in generative video models through Context Learning in Video Generation. The benchmark comprises 6 categories and 47 subcategories covering diverse reasoning-oriented video generation tasks, together with an Adaptive Video Evaluator for scalable and reliable assessment.

Experiments reveal that current video models still exhibit substantial limitations in multimodal reasoning, particularly in physical simulation, logical reasoning, and multi-turn interaction. By identifying these bottlenecks, CLVG-Bench provides a comprehensive diagnostic framework and highlights the need for deeper integration of visual understanding and generative reasoning toward more capable video foundation models.

Acknowledgements

This work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant no. 2025C01128), and the ZJU-Angelalign R&D Center for Intelligence Healthcare.

References

1. Agrawal, L.A., Tan, S., Soylu, D., Ziems, N., Khare, R., Opsahl-Ong, K., Singhvi, A., Shandilya, H., Ryan, M.J., Jiang, M., Potts, C., Sen, K., Dimakis, A.G., Stoica, I., Klein, D., Zaharia, M., Khattab, O.: Gepa: Reflective prompt evolution can outperform reinforcement learning (2025), https://arxiv.org/abs/2507.19457 9, 10
2. Andrew, J.D.: Concepts in film theory. Oxford University Press (1984) 6
3. Bansal, H., Lin, Z., Xie, T., Zong, Z., Yarom, M., Bitton, Y., Jiang, C., Sun, Y., Chang, K.W., Grover, A.: Videophy: Evaluating physical commonsense for video generation (2024), https://arxiv.org/abs/2406.03520 3
4. Bansal, H., Peng, C., Bitton, Y., Goldenberg, R., Grover, A., Chang, K.W.: Videophy-2: A challenging action-centric physical commonsense evaluation in video generation (2025), https://arxiv.org/abs/2503.06800 3
5. Bordwell, D., Thompson, K., Smith, J.: Film art: An introduction, vol. 7. McGraw-Hill New York (2008) 6
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 1
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 4
8. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 4
9. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: Flownet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015) 3
10. Eisenstein, J., Nagpal, C., Agarwal, A., Beirami, A., D'Amour, A., Dvijotham, D., Fisch, A., Heller, K., Pfohl, S., Ramachandran, D., Shaw, P., Berant, J.: Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking (2024), https://arxiv.org/abs/2312.09244 3
11. Google: Gemini 3 pro model card. Tech. rep., Google DeepMind (12 2025), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, model Release: November 2025 6
12. Google: Veo 3 launch. https://cloud.google.com/blog/products/ai-machine-learning/veo-3-fast-available-for-everyone-on-vertex-ai (2025), accessed: March 3, 2026 4, 9, 11
13. Guo, M.H., Chu, X., Yang, Q., Mo, Z.H., Shen, Y., Li, P.l., Lin, X., Zhang, J., Chen, X.S., Zhang, Y., et al.: Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. arXiv preprint arXiv:2505.16770 (2025) 4, 5

14. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 9175–9184 (June 2026) 4
15. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., Richardson, E., Shiran, G., Chachy, I., Chetboun, J., Finkelson, M., Kupchick, M., Zabari, N., Guetta, N., Kotler, N., Bibi, O., Gordon, O., Panet, P., Benita, R., Armon, S., Kulikov, V., Inger, Y., Shiftan, Y., Melumian, Z., Farbman, Z.: Ltx-2: Efficient joint audio-visual foundation model (2026), https://arxiv.org/abs/2601.03233 9
16. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026) 4, 11
17. Han, H., Li, S., Chen, J., Yuan, Y., Wu, Y., Leong, C.T., Du, H., Fu, J., Li, Y., Zhang, J., Zhang, C., jia Li, L., Ni, Y.: Video-bench: Human-aligned video generation benchmark (2025), https://arxiv.org/abs/2504.04907 3, 5
18. He, H., Wang, J., Zhang, J., Xue, Z., Bu, X., Yang, Q., Wen, S., Xie, L.: Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. arXiv preprint arXiv:2512.07826 (2025) 4, 5
19. He, X., Jiang, D., Nie, P., Liu, M., Jiang, Z., Su, M., Ma, W., Lin, J., Ye, C., Lu, Y., Wu, K., Schneider, B., Do, Q.D., Li, Z., Jia, Y., Zhang, Y., Cheng, G., Wang, H., Zhou, W., Lin, Q., Zhang, Y., Zhang, G., Huang, W., Chen, W.: Videoscore2: Think before you score in generative video evaluation (2025), https://arxiv.org/abs/2509.22799 3
20. He, X., Jiang, D., Zhang, G., Ku, M., Soni, A., Siu, S., Chen, H., Chandra, A., Jiang, Z., Arulraj, A., Wang, K., Do, Q.D., Ni, Y., Lyu, B., Narsupalli, Y., Fan, R., Lyu, Z., Lin, Y., Chen, W.: Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. ArXiv **abs/2406.15252** (2024), https://arxiv.org/abs/2406.15252 3, 5
21. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning (2022), https://arxiv.org/abs/2104.08718 4
22. Hu, H., Chan, K.C., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al.: Instruct-imagen: Image generation with multi-modal instruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4754–4763 (2024) 2
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3, 4
24. 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) 2
25. 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) 4, 5

26. Ji, P., Xiao, C., Tai, H., Huo, M.: T2vbench: Benchmarking temporal dynamics for text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5325–5335 (2024) 3, 4
27. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025) 1
28. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In: Findings of the association for computational linguistics: EMNLP 2024. pp. 3843–3860 (2024) 1
29. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025) 2, 4
30. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025) 5
31. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 4
32. Ju, X., Ye, W., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Xu, Q.: Fulldit: Multi-task video generative foundation model with full attention. arXiv preprint arXiv:2503.19907 (2025) 2
33. 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) 2, 11
34. 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) 4
35. Ku, M., Wei, C., Ren, W., Yang, H., Chen, W.: Anyv2v: A tuning-free framework for any video-to-video editing tasks. arXiv preprint arXiv:2403.14468 (2024) 2
36. Li, M., Xie, C., Wu, Y., Zhang, L., Wang, M.: Five-bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16672–16681 (2025) 5
37. Liao, C., Liu, L., Wang, X., Luo, Z., Zhang, X., Zhao, W., Wu, J., Li, L., Tian, Z., Huang, W.: Mogao: An omni foundation model for interleaved multi-modal generation. arXiv preprint arXiv:2505.05472 (2025) 4
38. Liu, S.: Zero-shot voice conversion with diffusion transformers. arXiv preprint arXiv:2411.09943 (2024) 7
39. Liu, Y., Cun, X., Liu, X., Wang, X., Zhang, Y., Chen, H., Liu, Y., Zeng, T., Chan, R., Shan, Y.: Evalcrafter: Benchmarking and evaluating large video generation models. arXiv preprint arXiv:2310.11440 (2023) 3, 4
40. Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., Hou, L.: Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation (2023), https://arxiv.org/abs/2311.01813 3, 4
41. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024) 4

42. Meng, F., Liao, J., Tan, X., Shao, W., Lu, Q., Zhang, K., Cheng, Y., Li, D., Qiao, Y., Luo, P.: Towards world simulator: Crafting physical commonsense-based benchmark for video generation. arXiv preprint arXiv:2410.05363 (2024) 7
43. OpenAI: Sora: A video generation model. https://openai.com/zh-Hans-CN/index/sora-2/ (2025), accessed: March 4 2026 9, 11
44. PixVerse: Pixverse: AI-powered image and video editing platform. https://app.pixverse.ai/ (2023), accessed: March 3, 2026 4
45. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024) 2
46. Ren, Z., Wei, Y., Yu, X., Luo, G., Zhao, Y., Kang, B., Feng, J., Jin, X.: Videoworld 2: Learning transferable knowledge from real-world videos. arXiv preprint arXiv:2602.10102 (2026) 4
47. Ren, Z., Wei, Y., Yu, X., Luo, G., Zhao, Y., Kang, B., Feng, J., Jin, X.: Videoworld 2: Learning transferable knowledge from real-world videos. arXiv preprint arXiv:2602.10102 (2026) 5, 7
48. Seed, B.: Seed2.0 model card: Towards intelligence frontier for real-world complexity. Tech. rep., 2026a. Technical Report (2026) 4, 6, 9, 11
49. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025) 2, 9, 11
50. Sun, K., Huang, K., Liu, X., Wu, Y., Xu, Z., Li, Z., Liu, X.: T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8406–8416 (2025) 3, 4
51. Sun, S., Liang, X., Fan, S., Gao, W., Gao, W.: Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7105–7113 (2025) 4, 5
52. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024) 4
53. Team, K., Chen, J., Ci, Y., Du, X., Feng, Z., Gai, K., Guo, S., Han, F., He, J., He, K., et al.: Kling-omni technical report. arXiv preprint arXiv:2512.16776 (2025) 2, 4
54. Team, T.H.F.M.: Hunyuanvideo 1.5 technical report (2025), https://arxiv.org/abs/2511.18870 9
55. Vidu: Vidu: AI-powered video generation platform. https://www.vidu.cn/ (2024), accessed: 2026-03-03 4
56. 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) 2, 4, 9, 11
57. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023) 4
58. Wang, M., Wang, R., Lin, J., Ji, R., Wiedemer, T., Gao, Q., Luo, D., Qian, Y., Huang, L., Hong, Z., et al.: A very big video reasoning suite. arXiv preprint arXiv:2602.20159 (2026) 4
59. Wang, M., Wang, R., Lin, J., Ji, R., Wiedemer, T., Gao, Q., Luo, D., Qian, Y., Huang, L., Hong, Z., et al.: A very big video reasoning suite. arXiv preprint arXiv:2602.20159 (2026) 5, 7
60. Wang, S., Pei, M., Sun, L., Deng, C., Li, Y., Shao, K., Tian, Z., Zhang, H., Wang, J.: Spatialviz-bench: A cognitively-grounded benchmark for diagnosing spatial vi-

sualization in mllms. In: The Fourteenth International Conference on Learning Representations (2025) 4
61. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024) 4
62. Wang, Y., Liu, J., Gao, S., Feng, B., Tang, Z., Gai, X., Wu, J., Liu, Z.: V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 658–668. Springer (2025) 2
63. Wei, C., Liu, Q., Ye, Z., Wang, Q., Wang, X., Wan, P., Gai, K., Chen, W.: Univideo: Unified understanding, generation, and editing for videos. arXiv preprint arXiv:2510.08377 (2025) 2, 4, 5, 9, 11
64. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al.: Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 (2022) 1
65. Wei, J., Zhang, X., Li, Y., Wang, Y., Zhang, Y., Chen, Z., Tang, Z., Xu, W., Liu, Z.: Univbench: Towards unified evaluation for video foundation models (2026), https://arxiv.org/abs/2602.21835 3, 5, 7
66. 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) 2, 5
67. Wu, J.Z., Fang, G., Fu, D.J., Kanakagiri, V.A.R., Iandola, F., Keutzer, K., Hsu, W., Dong, Z., Shou, M.Z.: Veditbench: Holistic benchmark for text-guided video editing (2026), https://arxiv.org/abs/2602.21835 3
68. Wu, J., Zhang, X., Yuan, H., Zhang, X., Huang, T., He, C., Deng, C., Zhang, R., Wu, Y., Long, M.: Visual generation unlocks human-like reasoning through multimodal world models. arXiv preprint arXiv:2601.19834 (2026) 7
69. Wu, R., Chen, L., Yang, T., Guo, C., Li, C., Zhang, X.: Lamp: Learn a motion pattern for few-shot video generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7089–7098 (2024) 4
70. Xu, J., Huang, Y., Cheng, J., Yang, Y., Xu, J., Wang, Y., Duan, W., Yang, S., Jin, Q., Li, S., Teng, J., Yang, Z., Zheng, W., Liu, X., Ding, M., Zhang, X., Gu, X., Huang, S., Huang, M., Tang, J., Dong, Y.: Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation (2024), https://arxiv.org/abs/2412.21059 3
71. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5288–5296 (2016) 4
72. Yang, Y., Fan, K., Sun, S., Li, H., Zeng, A., Han, F., Zhai, W., Liu, W., Cao, Y., Zha, Z.J.: Videogen-eval: Agent-based system for video generation evaluation (2025), https://arxiv.org/abs/2503.23452 3, 5
73. 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) 9
74. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025) 2
75. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025) 4

76. Ye, Z., He, X., Liu, Q., Wang, Q., Wang, X., Wan, P., Zhang, D., Gai, K., Chen, Q., Luo, W.: Unic: Unified in-context video editing. arXiv preprint arXiv:2506.04216 (2025) 5
77. Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Huang, Z., Guestrin, C., Zou, J.: Textgrad: Automatic "differentiation" via text (2024), https://arxiv.org/abs/2406.07496 9
78. Zhang, P., Jia, Z., Liu, K., Weng, S., Li, S., Shi, B.: Stage: Storyboard-anchored generation for cinematic multi-shot narrative. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 659–669 (2026) 2
79. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 3
80. Zhao, W., Han, Y., Tang, J., Wang, K., Song, Y., Huang, G., Wang, F., You, Y.: Dynamic diffusion transformer. arXiv preprint arXiv:2410.03456 (2024) 4