RAVEN-EVAL: RUBRIC-GUIDED AUTOMATIC EVALUATION FOR AI VIDEO GENERATION MODELS BASED ON LMM PREFERENCE JUDGEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

AI video generation has entered widespread commercial use. As a result, quality differences among videos produced by state-of-the-art AI video generation models (AIVGMs) have become increasingly difficult to discern using conventional evaluation criteria, such as visual fidelity and semantic instruction following. Meanwhile, human evaluation now requires more expertise and sustained attention, substantially increasing annotation costs. This calls for automated evaluation that can reliably distinguish fine-grained differences among advanced AIVGMs with minimal human intervention. To address this challenge, we present **RAVEN-Eval**, a rubric-guided automated evaluation framework for AIVGMs, built primarily on the LMM-as-a-judge paradigm. Through an automatic task curation and quality-filtering pipeline, RAVEN-Eval curates 150 text-to-video (T2V) tasks and 100 image-to-video (I2V) tasks, and systematically collects more than $4,500$ AIGVs. At its core, RAVEN-Eval adopts **rubric-guided automated LMM preference judgment**, in which LMM judges conduct pairwise comparisons according to task-specific rubrics. It further introduces an **anchor-based model insertion approach** to reduce the evaluation cost of incorporating new models. Finally, we evaluate 20 high-performance AIVGMs, as well as the judging capabilities of 13 LMM judges, and establish the **RAVEN-Eval Leaderboards**. Overall, RAVEN-Eval paves a scalable path for automatic evaluation of rapidly evolving AIVGMs.

1 INTRODUCTION

With the widespread commercialization of AI-generated videos (AIGVs), **reliable** and **scalable** evaluation of AI video generation models (AIVGMs) has become increasingly critical since effective benchmarks guide model selection and characterize model strengths across diverse generation scenarios. However, the rapid progress of state-of-the-art (SOTA) AIVGMs is outpacing conventional evaluation paradigms, creating new demands for more discriminative assessment. Here, we identify **two major challenges**.

First, widely used leaderboards such as (Artificial Analysis, 2026) rely heavily on human preference annotations. As the number of AIGVs grows, the associated cost becomes increasingly substantial. Moreover, improvements in AIVGMs make quality differences more subtle, requiring finer-grained perception. Annotators must therefore invest greater expertise, attention, and effort to make reliable decisions, further increasing cost and potentially reducing annotation consistency.

Second, although semantic instruction following, basic physical consistency, and visual fidelity remain essential evaluation criteria, recent SOTA AIVGMs have become increasingly competitive along these dimensions. Consequently, frameworks based on fixed dimensions or handcrafted metrics (Huang et al., 2024; Wang et al., 2025) are less effective at distinguishing advanced models. Meanwhile, commercialization raises user expectations for precise and professional generation, making the remaining capability gaps more evident in **fine-grained detail control** and **general intelligence** under **task-dependent** and **context-specific** scenarios. Capturing these differences therefore requires **task-specific, fine-grained preference judgments**. This motivates our central question: *How can an automatic evaluation framework with limited human involvement faithfully capture the capability differences among recent SOTA AIVGMs?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this, we propose **RAVEN-Eval**, a task-centric rubric-guided automatic evaluation framework that employs large multimodal models (LMMs) for preference judgement, as illustrated in Fig. 1. Its feasibility is supported by **two observations**. **First**, current LMMs exhibit strong video understanding and can reliably identify fine-grained differences under effective rubric guidance. Compared with human annotators, they are less susceptible to irrelevant visual content or contextual distractions, enabling more consistent attention to task-relevant details. **Second**, absolute scoring may fail to distinguish high-performing AIVGMs because of limited score granularity. Pairwise preference judgments offer a more suitable alternative, as relative comparison places lower perceptual and decision-making demands on the evaluator than precise absolute scoring. Our main contributions are summarized as follows:

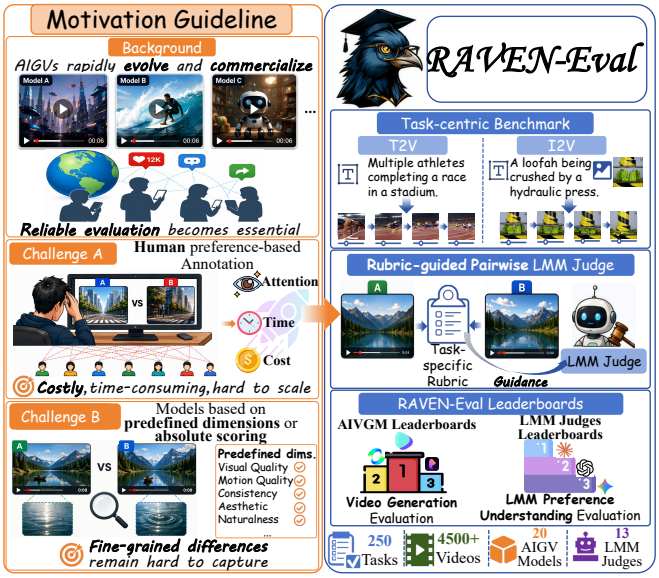


Figure 1: To address the rising cost and limited fine-grained discrimination of existing AIVGM evaluation methods, we introduce **RAVEN-Eval**, a task-centric, rubric-guided LMM pairwise judgment framework for T2V and I2V tasks.

- We construct the **RAVEN-Eval Benchmark** for fine-grained detail control and task-dependent model intelligence. Rather than relying on conventional predefined dimensions, we build task-centric evaluation scenarios with context-specific requirements through an LMM-assisted generation and quality-filtering pipeline. The benchmark comprises 250 diverse T2V and I2V tasks and 4,669 high-quality generated videos.
- We introduce **rubric-guided pairwise preference judgement** under the **LMM-as-a-judge** paradigm. Each task is equipped with a customized rubric that guides LMM judges to compare details and determine pairwise preferences. Model capabilities are estimated through tie-aware optimization, while an anchor-based model insertion trick further enables efficient incorporation of new AIVGMs.
- We establish two complementary **RAVEN-Eval Leaderboards**: the **AIGV Leaderboards** for ranking 20 recent open-source and proprietary generation models, and the **Judge Leaderboards** for benchmarking 13 general-purpose LMM judges against human preferences. Their strong performance across diverse general LMMs demonstrates that RAVEN-Eval has the potential to be a **LMM-judge-agnostic evaluation harness** for LMM judges.

2 Related Works

Existing AIVGM evaluation can be roughly categorized into **human annotation-based** platforms and **automatic frameworks** using trained models or predefined toolkits.

Arena-style platforms such as *Artificial Analysis* (Artificial Analysis, 2026), *OpenCompass* (Cao et al., 2026), and *Video Arena* (Video Arena, 2026) typically adopt human pairwise comparison protocols and dynamic scheduling systems, such as Elo-style ranking (Dangauthier et al., 2007; Glickman, 1999), to continuously update model scores from human votes.

Recent evaluation approaches mainly rely on pretrained models and handcrafted metrics. *EvalCrafter* (Liu et al., 2024) and the *VBench* series (Huang et al., 2024) establish benchmarks and toolkits based on predefined dimensions and corresponding automatic metrics. *VBench++* (Huang et al., 2025) and *VBench-2.0* (Zheng et al., 2025) further broaden the capability coverage and granularity of this dimension-based paradigm. *VideoScore* (He et al., 2024), *Video-Bench* (Han et al., 2025)

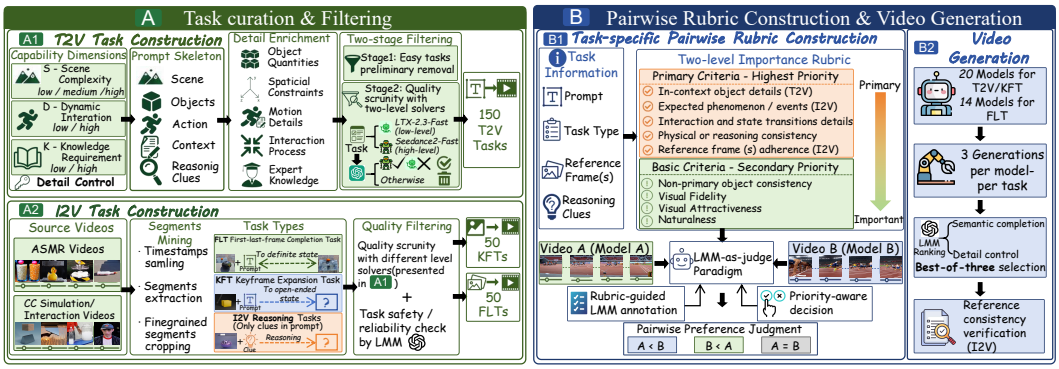


Figure 2: The pipelines of the task curation, rubric generation and video generation.

explores using vision-language models as automatic evaluators. *T2V-CompBench* (Sun et al., 2025) evaluates compositional T2V generation through MLLM-, detection-, and tracking-based metrics tailored to different compositional categories, but still relies on absolute scores. *VideoPhy* (Bansal et al., 2025a) and *VideoPhy-2* (Bansal et al., 2025b) focus on physical commonsense through tasks that test real-world consistency. Scorers such as *AIGV-Assessor* (Wang et al., 2025) and *LOVE* (Wang et al., 2026) investigate LMM-based AIGV assessment and mean opinion score (MOS) prediction. Despite their effectiveness, existing approaches still have limitations in evaluating SOTA AIVGMs as described above. Human evaluation requires substantial resources, while most automatic methods rely on predefined dimensions or absolute scoring, which may not sufficiently capture subtle capability differences among advanced models. These observations also provide motivation for our work.

3 The RAVEN-Eval Benchmark

3.1 Task Curation

To evaluate AIVGMs across diverse scenarios, we construct three task types: **T2V generation**, **keyframe expansion task** (**KFT**), and **first–last frame completion task** (**FLT**). Each is generated and filtered through a dedicated automatic pipeline, as illustrated in Fig. 2, with dataset statistics reported in Fig. 3. Examples are provided in *Supplementary Materials (Supp.) Sec. A.6*.

T2V Tasks As discussed above, evaluating SOTA AIVGMs requires greater emphasis on **fine-grained detail control** and **model intelligence**. We therefore define three complementary **capability dimensions** for prompt construction: **scene complexity (S)**, **dynamic interaction complexity (D)**, and **expert-level knowledge requirement (K)**. S controls the difficulty of object and spatial-detail realization. **S0** scenes contain only a few objects with simple layouts; **S1** scenes contain multiple distinct objects with nontrivial compositional relationships; and **S2** scenes contain groups of objects arranged in complex spatial structures. D controls motion and interaction complexity. **D0** tasks involve nearly static objects or limited motion with simple interactions, whereas **D1** tasks require coordinated, highly dynamic, and complex interaction processes. K introduces domain-specific knowledge and reasoning requirements. **K0** tasks depict everyday situations requiring no additional expertise, while **K1** tasks require accurate representation of specialized phenomena, processes, or constraints, including physical experiments, chemical reactions, material transformations, precise text rendering, specialized photographic or aesthetic styles, and other knowledge-intensive scenarios. We additionally designate approximately 20% of the tasks as **reasoning-based tasks** to further assess general reasoning ability. Combining S, D, and K enables each task to jointly evaluate multiple advanced capabilities rather than a single predefined quality dimension.

For each valid task, we use the proprietary LMM *GPT-5.5* (OpenAI, 2026c) to generate a **prompt skeleton** defining the core scene and interactions according to its assigned capability dimensions, and then **enrich** it with task-specific details. For **S1** and **S2** tasks, we add fine-grained compositional and spatial constraints together with explicit object-count requirements. For **D1** tasks, we provide more detailed specifications of motion patterns and interaction processes. For **K1** tasks, we incorporate

domain-specific descriptions of the expected phenomena or processes. For **reasoning-based tasks**, however, the prompt contains only **clues**, without explicitly stating the expected visual outcome. The intended phenomenon or event, which is subsequently included in the task rubric, must be **uniquely inferable** from the provided clues using general reasoning and relevant domain knowledge.

Following the scheme above, we construct a large candidate pool with balanced coverage across combinations of S, D, and K. We then apply a **two-stage automatic filtering** pipeline to retain informative T2V tasks. First, we remove all **S0** and **S1-D0** tasks. These low-difficulty combinations are generated during prompt generation to preserve a complete difficulty hierarchy, but are unlikely to reveal meaningful differences among advanced AIVGMs and are therefore excluded from the final benchmark. In the second stage, inspired by the data-curation paradigm with **two-level solvers** in (Kulikov et al., 2026), we select *Seedance2 Fast* (Seedance Team et al., 2026) and *LTX 2.3 Fast* (LTX-2.3, 2026) as preliminary **solvers** with different capability levels. The former is a high-performance proprietary model, whereas the latter is a smaller open-weight model. Both solvers generate videos for every task retained after the first-stage filtering. We then use *GPT-5.4-mini* (OpenAI, 2026b) to assess prompt fulfillment on a five-point ordinal scale, jointly considering core semantic completion and realization of specified details. We retain tasks for which *Seedance2 Fast* scores at least as high as *LTX 2.3 Fast*. This consistency criterion aims to reduce the influence of anomalous tasks without imposing absolute score thresholds, thereby preventing prior assumptions from exerting undue influence on task selection.

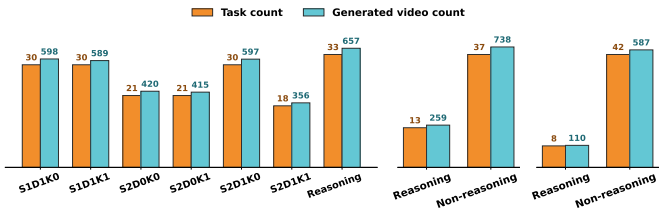


Figure 3: Statistical data of the **RAVEN-Eval Benchmark**. It comprises 250 tasks and 4,669 videos in total.

I2V Tasks We adopt a separate construction strategy for FLTs and KFTs. Since reference images impose stronger visual constraints than text alone, these tasks are better suited to evaluating whether models can generate **well-specified, domain-relevant interaction and transformation processes**.

We collect *Creative Commons-licensed* (*CC*) videos, with a focus on **autonomous sensory meridian response** (**ASMR**) content (Barratt & Davis, 2015). Such videos typically capture brief yet information-dense processes in specialized settings, such as hydraulic presses crushing materials or red-hot iron balls penetrating different substances. Their reliance on physical and material behavior, together with durations well suited to mainstream AIVGMs, makes them valuable sources for challenging I2V tasks. We collect approximately 1,000 ASMR videos from *YouTube* and about 500 additional videos featuring complex object interactions. In the final benchmark, the KFT set contains 31 tasks derived from ASMR videos and 19 from other sources, while the FLT set contains 35 ASMR-derived tasks and 15 other tasks.

For each source video, we sample frames every 3 seconds for *GPT-5.4-mini*-based **segment mining**, identifying complete transformations or interactions and estimating their temporal boundaries, then refine these boundaries using 16 uniformly sampled frames per interval. Pooled pixel-vector similarity filters out segments with insufficient first-to-last-frame variation. For each retained segment, *GPT-5.4-mini* generates an I2V prompt. **FLT** prompts specify coherent transitions with endpoints strictly matching the reference frames and intermediate processes reasonably inferred from the source segment. **KFT** prompts align the initial state with the reference keyframe while allowing open-ended, logically consistent subsequent events. Approximately 20% of I2V tasks require reasoning: prompts provide reference-image-derived clues that uniquely imply a predictable phenomenon or event without revealing the outcome, requiring AIVGMs to infer and depict it. Following the Stage-2 T2V protocol, *Seedance2 Fast*, *LTX 2.3 Fast*, and *GPT-5.4-mini* filter the tasks. A final GPT-assisted safety and reliability check removes potentially harmful or copyright-restricted content.

Rubric Generation As discussed above, recent LMMs are suitable for **fine-grained multimodal understanding tasks** when provided with clear checklists. Therefore, constructing **task-centric rubrics** to guide LMM judges to distinguish the quality differences between videos is critical. We also visualize this process in Fig. 2. For each task, we construct a two-level rubric comprising

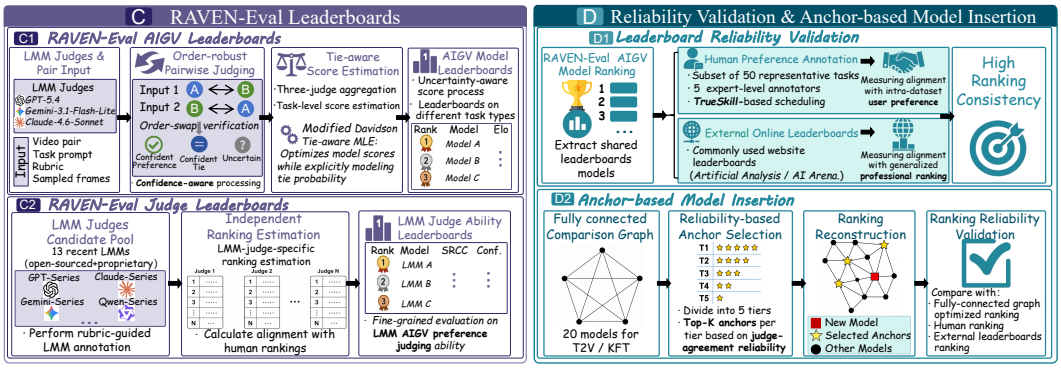


Figure 4: The pipelines of the subsequent processes.

task-specific primary criteria and **shared basic criteria**. The primary criteria assess the realization of task-defining details. The basic criteria assess foundational aspects, including global semantic alignment, non-primary object consistency, visual fidelity, aesthetic appeal, and naturalness. Both levels contribute to the overall preference judgment. The priority assigned to task-specific criteria directs the judge's attention toward details that distinguish otherwise competitive outputs; it does not automatically override substantial semantic or perceptual failures. When two videos satisfy the basic requirements to a comparable degree, task-specific detail realization becomes the principal basis for distinguishing them. Substantial failures in fundamental aspects still remain relevant. The rubric therefore supports a **comprehensive assessment** of task-conditioned video quality while making the criteria most relevant to the task explicit.

For T2V tasks, we use *GPT-5.5* to extract the enriched details previously added to the prompts and combine them with the original prompt content to form *itemized evaluation criteria*. For I2V tasks, since the reference images provide richer semantic information and the prompt text mainly details the required events or phenomena, we directly reuse the prompt content as the primary criteria. Furthermore, for all the reasoning-based tasks, since the prompts only contain clues, we supplement the primary compliance criteria according to the task construction records, providing detailed requirements for the expected content presentation. The **basic criteria** are assigned lower priority than the primary criteria, and the same set is applied uniformly to all tasks. They assess non-primary object consistency, visual fidelity, aesthetic appeal, and naturalness. Detailed definitions are provided in *Supp. Sec. A.3*.

Benchmark Video Generation We demonstrate the video generation process in Fig. 2. Since video generation models differ in their capable task ranges, we evaluate 20 models on both T2V tasks and KFTs, and 14 models on FLTs. For a clearer presentation, these models are reported together in the leaderboard shown in Tab. 1. For each task, we randomly select a configuration from the combinations of aspect ratio, resolution, and duration commonly supported by the models. If an individual model does not support the selected configuration, we use its closest available configuration instead. All videos are generated without audio.

To reduce the effect of randomness in generation, each model generates 3 videos for each task. We use *GPT-5.4-mini* to rank the candidates based on task-semantic completion and detail control, and retain the highest-ranked video. For KFTs and FLTs, we further verify consistency with the reference images. Specifically, we compare the first frame of each KFT video, or the first and last frames of each FLT video, with the corresponding reference images using the similarity between pooled pixel-value vectors. Generated videos with a similarity below the threshold of 0.98 are excluded.

4 RAVEN-EVAL LEADERBOARDS

RAVEN-Eval AIVGM Leaderboards We perform LMM-as-a-judge preference judgement for each task using three LMMs: *GPT-5.4* (OpenAI, 2026a), *Claude Sonnet 4.6* (Anthropic, 2026c), and *Gemini 3.1 Flash-Lite* (DeepMind, 2026a). To avoid selection bias, these judges are chosen **a priori** based on the trade-off between their recognized multimodal understanding capabilities and inference

costs. RAVEN-Eval is designed as a **LMM-judge-agnostic evaluation harness** that accommodates heterogeneous LMM judges. To support multiple judges, we also adopt a trade-off between model harnessing and input granularity: videos are submitted directly to LMM judges supporting native video input; otherwise, frames are sampled at one frame per second and provided in chronological order using the required input format. For every comparison, the identities of the AIVGMs are **concealed from the judge** to prevent bias. Detailed settings are provided in *Supp. Sec. A.5*.

Because some AIGV pairs remain difficult to distinguish with task-specific rubrics, judges may output a "tie". To control judgment stability, we adopt an **order-robust pairwise judging and confidence-aware processing** protocol, under which each pair is evaluated twice with reversed presentation orders. Both outputs are first mapped back to the corresponding model identities. A preference that remains consistent after reversal, or two "tie" outputs, is treated as a **confident judgment**; all other cases, including an order-inconsistent preference or a single "tie", are treated as **uncertain judgments** and downweighted during capability estimation. For each task, all feasible model pairs are evaluated by all judges, forming a **fully connected comparison graph** that also supports the subsequent anchor-based insertion experiments. We then estimate task-level model capabilities using a modified *Davidson tie-aware maximum likelihood estimation objective* (Davidson, 1970):

$$
\begin{aligned}
&\arg\max_{\{s_i\}_{i=1}^{N},\,\gamma} \frac{1}{|\mathcal{P}|}\sum_{p=(i,j)\in\mathcal{P}} \lambda_p \frac{1}{M}\sum_{m=1}^{M}\Big[w_{p,i>j}^{(m)}\log P(i\succ j)+\\
&w_{p,j>i}^{(m)}\log P(j\succ i)+w_{p,i\sim j}^{(m)}\log P(i\sim j)\Big]-\frac{\alpha}{N}\sum_{i=1}^{N}s_i^2-\beta\gamma^2,\\
&\text{s.t.}\quad \sum_{i=1}^{N} s_i=0,\quad \theta_i=\exp(s_i),\quad \eta=\exp(\gamma),\\
&P(i\succ j)=\frac{\theta_i}{\theta_i+\theta_j+\eta\sqrt{\theta_i\theta_j}},\\
&P(j\succ i)=\frac{\theta_j}{\theta_i+\theta_j+\eta\sqrt{\theta_i\theta_j}},\\
&P(i\sim j)=\frac{\eta\sqrt{\theta_i\theta_j}}{\theta_i+\theta_j+\eta\sqrt{\theta_i\theta_j}}.
\end{aligned}
\tag{1}
$$

Here, $\mathcal{P}$ denotes the set of valid pairwise comparisons for a task, where $p=(i,j)$ represents a comparison between two AIVGMs. N and M denote the numbers of AIVGMs and LMM judges, respectively. The variable $s_i\in\mathbb{R}$ is the latent capability score of model i, while $\gamma\in\mathbb{R}$ controls the overall tie propensity through $\theta_i=\exp(s_i)$ and $\eta=\exp(\gamma)$. Let $c_p^{(m)}\in\{0,1\}$ indicate whether the m-th judge provides a confident judgment for pair p. The pair-level reliability weight is defined as the proportion of confident judgments, $\lambda_p=\frac{1}{M}\sum_{m=1}^{M}c_p^{(m)}=1-\ell_p/M$, where $\ell_p=\sum_{m=1}^{M}(1-c_p^{(m)})$ is the number of uncertain judgments. The outcome weights $w_{p,i>j}^{(m)}$, $w_{p,j>i}^{(m)}$, and $w_{p,i\sim j}^{(m)}$ encode the result provided by the m-th judge. For a confident judgment, the weight of the resolved outcome is set to 1, while the other two weights are set to 0. For an uncertain judgment, we assign $w_{p,i>j}^{(m)}=w_{p,j>i}^{(m)}=1/2$ and $w_{p,i\sim j}^{(m)}=0$, representing equal uncertainty over the two preference directions.

The coefficients α and β regularize the latent capability scores and the tie parameter, respectively. The constraint $\sum_{i=1}^{N}s_i=0$ resolves the location non-identifiability of the latent scores. We optimize the objective independently for each task for 1,000 iterations. For each task category c, a model's final capability $\bar{s}_{i,c}$ is calculated as its mean latent score across all tasks in that category minus the standard deviation of these scores. The resulting capability scores are then linearly rescaled to an Elo-style range ($\mathrm{Elo}_{i,c}=1000+\frac{400}{\ln 10}\bar{s}_{i,c}$) for intuitive presentation in Tab. 1. The process of building the leaderboard is shown in Fig. 4.

Reliability Verification To verify the reliability of the *RAVEN-Eval AIVGM Leaderboards*, we conduct human preference experiments on 50 tasks, including 30 T2V tasks, 10 KFTs, and 10 FLTs. T2V tasks are sampled to preserve the original distribution of the S, D, and K levels, while I2V

Category	T2V				FLT				KFT			
AIVGM	*Rank*	*Score*	*CI*	*# Votes*	*Rank*	*Score*	*CI*	*# Votes*	*Rank*	*Score*	*CI*	*# Votes*
Seedance 2	1	**840.7**	±21	8487	1	**890.5**	±10	1941	1	**1094.3**	±16	**2847**
G-I-V (Grok Imagine Video, 2026)	2	*791.5*	±19	8475	–	–	–	–	3	978.9	±16	*2844*
Kling 3.0 Pro (Kling 3.0, 2026)	3	764.6	±23	8496	4	817.5	±20	1935	4	839.1	±19	2832
HappyHorse V1.0	4	762.6	±28	8487	–	–	–	–	10	690.1	±27	2829
Kling 3.0 Omni	5	733.6	±21	8487	2	*854.7*	±12	1935	8	720.3	±20	2826
Seedance 2 Fast	6	727.7	±15	8508	5	799.5	±16	**1947**	5	828.4	±30	2835
Veo 3.1 Fast (Veo 3.1, 2025)	7	680.0	±13	8484	12	130.9	±35	*1944*	9	706.8	±17	2841
Veo 3.1 Lite	8	594.6	±21	8499	10	363.3	±30	*1944*	12	641.6	±16	2823
PixVerse C1 (PixVerse C1, 2026)	9	567.7	±16	8517	7	749.7	±14	1938	2	*1047.4*	±14	2832
PixVerse V6 (PixVerse V6, 2026)	10	555.3	±17	8490	8	609.5	±35	1932	11	686.6	±28	2838
Kling 3.0 Std	11	536.2	±25	8484	6	786.7	±13	*1944*	6	785.1	±18	2832
Wan 2.7 (WanVideo, 2026)	12	509.9	±23	8496	3	848.4	±13	1935	7	731.8	±14	2832
PixVerse V5.5 (PixVerse V5.5, 2025)	13	470.8	±15	8493	11	274.0	±28	1935	16	328.5	±26	2832
Seedance 1.5 Pro (Chen et al., 2025)	14	393.9	±16	*8529*	9	506.7	±21	1935	15	337.9	±29	2835
LTX 2.3 Fast	15	344.1	±23	8520	14	-468.1	±14	*1944*	17	-155.9	±23	2841
Wan 2.6	16	342.2	±19	8484	–	–	–	–	13	638.3	±22	2826
LTX 2.3 Pro (LTX-2.3, 2026)	17	13.3	±25	**8538**	13	73.5	±15	1932	18	-203.0	±22	2838
Hailuo 2.3 (Hailuo2.3, 2025)	18	5.8	±27	8520	–	–	–	–	14	400.1	±19	2826
LTX 2.0 Pro (HaCohen et al., 2026)	19	-27.6	±23	8493	–	–	–	–	19	-323.2	±19	2832
LTX 2.0 Fast	20	-162.2	±19	8514	–	–	–	–	20	-336.8	±22	*2844*

Table 1: The RAVEN-Eval AIVGM Leaderboards. # Votes denotes the number of effective LMM pairwise judgements involving the model. "CI" denotes the 95% confidence interval. [Per column, the highest value is shown in **bold**, and the second in *italics*.]

tasks approximately retain the original ratio of reasoning to non-reasoning tasks. We recruit 5 experts to provide pairwise preference annotations. To reduce annotation cost, we adopt a *TrueSkill*-based adaptive scheduling strategy (Herbrich et al., 2006).

For each task, all annotators begin with the same cold-start pairs, with every model participating in 4 randomly sampled comparisons. We assess inter-annotator agreement by pooling the cold-start annotations across all tasks and computing nominal *Krippendorff's* α (Krippendorff, 1970), treating the two directional preferences and the tie as three categorical outcomes. The resulting α of 0.714 indicates acceptable agreement. After cold start, each annotator maintains an independent state and receives a separately scheduled sequence of comparisons. For each expert, the task-level ranking is updated using Eq. 1 after every 10 newly annotated pairs. Annotation terminates when all pairwise *Spearman rank correlation coefficient* (*SRCC*) values among the three most recent rankings exceed 0.95, and the mean σ across models falls below 0.3. For each task, we average the independently optimized capability scores from all annotators to obtain a consensus human score for every model. These scores are then averaged across the selected tasks to construct a human-reference ranking for each task type. Further details are provided in *Supp. Sec. A.4*.

We compare the **RAVEN-Eval AIVGM Leaderboards** with the human rankings and two external online leaderboards, *Artificial Analysis* (*AA*) (Artificial Analysis, 2026) and *Arena AI* (*Arena.*) (Arena, 2026). Since both external platforms provide only I2V leaderboards, we merge the KFT and FLT results into a unified I2V ranking. For each external comparison, we retain only the models shared with our evaluation set and compute the *bootstrap-stabilized SRCC* (detailed in *Supp. Sec. A.5*). The overall validation process is illustrated in Fig. 4, and results are reported in the last row of Tab. 2 for readability. The strong agreement across multiple human-annotation-based references demonstrates the reliability of our evaluation framework.

The Key Role of Rubrics To assess the effectiveness of rubric-guided pairwise evaluation, we conduct an ablation study using the same LMM judges. In *no-rubric pairwise evaluation* (**NR-PW**), the judges select the preferred video using only the prompt and video, without access to the rubric. In *rubric-guided absolute scoring* (**RG-AS**), each video is independently assigned a continuous score from 0 to 5 under the original task-specific rubric, and model scores are averaged across tasks. Our *rubric-guided pairwise evaluation* (**RG-PW**) directly compares each video pair using the corresponding rubric. We additionally evaluate two representative trained, dimension-based absolute scorers, *VideoScore* (He et al., 2024) and *LOVE* (Wang et al., 2026), using their publicly released weights. For each task category, both models score every generated video along their predefined dimensions. Scores from each dimension are linearly normalized to a common scale across all evaluated videos and AIVGMs, averaged across dimensions for each video, and then averaged across

videos to obtain the final score of each AIVGM. The resulting rankings are compared with our human-reference ranking and external leaderboards, as reported in Tab. 2. The results show that absolute scoring, including both rubric-based (*RG-AS*) and trained predefined-dimension evaluators, produces substantially weaker ranking consistency than our setting. Moreover, our setting clearly outperforms *NR-PW*, demonstrating the importance of task-specific rubric guidance.

4.1 Anchor-Based New Model Insertion

Although LMM judges reduce costs, the number of pairwise comparisons still grows quadratically with the size of the model pool. Maintaining a fully connected comparison graph is impractical for continuously updating the model pool. Inspired by (Zhu et al., 2024), we introduce an anchor-based insertion strategy (also shown in Fig. 4), in which newly added models are compared only with a subset of anchors from the existing model pool.

Category	**T2V**			**I2V**			
Setting	*Human*	*AA*	*Arena.*	*H. (KFT)*	*H. (FLT)*	*AA*	*Arena.*
NR-PW	*0.754*	*0.733*	*0.767*	*0.852*	*0.785*	*0.802*	*0.677*
RG-AS	0.672	0.455	0.622	0.633	0.585	0.432	0.548
VideoScore	0.723	0.611	0.595	0.654	0.542	0.577	0.536
LOVE	0.711	0.643	0.734	0.625	0.588	0.607	0.575
Ours	**0.872**	**0.835**	**0.810**	**0.903**	**0.821**	**0.851**	**0.714**

Table 2: Ablation results on the effectiveness of rubrics. The term "H." is short for "Human".

Anchor Selection Since fully connected comparison graphs are available for all 20 AIVGMs on both **T2V tasks** and **KFTs**, we conduct the following experiments independently for these two task categories. In each trial, we randomly select 15 models to form an initial pool and treat the remaining 5 as newly inserted models. For each candidate model i in the initial pool, we construct four one-dimensional rank vectors for each task category. Each vector has the same length as the number of evaluated tasks, with each element recording the final rank of model i on one task. All ranks are derived exclusively from the fully connected comparison graph of the selected 15 models. The vector $\mathbf{r}_i^{(\text{all})}$ is obtained using all three LMM judges ($M = 3$), while $\{\mathbf{r}_i^{(m)}\}_{m=1}^{3}$ are obtained separately using only the m-th judge ($M = 1$). We measure the agreement between the m-th judge and the joint evaluation of model i as $\rho_i^{(m)} = \text{SRCC}\,(\mathbf{r}_i^{(m)}, \mathbf{r}_i^{(\text{all})})$. The mean agreement and cross-judge standard deviation are defined as $\bar{\rho}_i = \frac{1}{3}\sum_{m=1}^{3} \rho_i^{(m)}$ and $v_i = \text{Std}\,(\rho_i^{(1)}, \rho_i^{(2)}, \rho_i^{(3)})$, respectively. We define the anchor reliability score as $q_i = \bar{\rho}_i - \tau v_i$, where τ controls the penalty for cross-judge inconsistency. A larger q_i indicates stronger agreement with the joint evaluation and greater stability across judges, making model i a more reliable anchor candidate in the current task space.

Model Insertion Simulation For each random split, we first reconstruct a leaderboard for the 15 initial models from their fully connected comparison graph, following the same procedure used for the *RAVEN-Eval AIVGM Leaderboards*. This process is performed independently for each task category. We then divide the 15 models into five consecutive tiers of three models according to the reconstructed ranking. Under our anchor-selection strategy, the top-K models with the highest reliability scores are selected from each tier, yielding $5K$ anchors in total. We evaluate $K \in \{1, 2\}$, corresponding to 5 and 10 anchors, respectively.

Category	**T2V**			**KFT**		
Setting	*Full*	*Human*	*AA*	*Full*	*Human*	*AA*
$K = 1, M = 1$ *(Rand.)*	0.767	0.732	0.705	0.792	0.771	0.722
$K = 1, M = 2$ *(Rand.)*	0.792	0.741	0.730	0.808	0.782	0.731
$K = 1, M = 3$ *(Rand.)*	0.811	0.752	0.749	0.820	0.783	0.728
$K = 1, M = 3$ *(Ours)*	*0.834*	0.770	0.755	0.851	0.798	0.747
$K = 2, M = 1$ *(Rand.)*	0.818	0.789	0.768	0.856	0.801	0.772
$K = 2, M = 2$ *(Rand.)*	0.820	0.785	0.781	0.864	0.822	0.791
$K = 2, M = 3$ *(Rand.)*	0.832	*0.791*	*0.785*	*0.878*	*0.828*	*0.801*
$K = 2, M = 3$ *(Ours)*	**0.854**	**0.811**	**0.814**	**0.903**	**0.857**	**0.835**

Table 3: Ablation results under different K, M settings using either the proposed anchor-selection strategy or random selection. "Full" uses the ranking from the complete comparison graph as the reference. "Rand." denotes random pick.

The remaining 5 models are treated as newly inserted models. We retain the fully connected graph among the 15 initial models and add only comparisons between each inserted model and the selected anchors. Applying the optimization in Eq. 1 to this combined graph jointly estimates the capability scores of all 20 models and reconstructs the expanded leaderboard. As a baseline, we repeat the same procedure using an equal number of randomly selected anchors from each tier. We further study the effect of the number of LMM judges using $M \in \{1, 2, 3\}$. More judges may provide more diverse evidence for distinguishing inserted models from the anchors, thereby improving ranking

Category	T2V				FLT				KFT			
LMM	*Rank*	*SRCC*	*KRCC*	*Conf.*	*Rank*	*SRCC*	*KRCC*	*Conf.*	*Rank*	*SRCC*	*KRCC*	*Conf.*
Claude Opus 4.6 (Anthropic, 2026a)	1	**0.853**	**0.686**	*84.8%*	2	0.867	*0.725*	79.1%	3	0.789	0.643	**81.4%**
GPT-5.4 (OpenAI, 2026a)	2	0.827	*0.673*	78.9%	1	**0.890**	**0.765**	80.1%	1	**0.817**	**0.687**	72.3%
GPT-5.6 Sol (OpenAI, 2026d)	3	*0.837*	0.660	68.5%	5	0.837	0.660	70.2%	4	0.776	0.606	72.3%
Claude Opus 4.8 (Anthropic, 2026b)	4	0.833	0.660	**85.3%**	3	*0.870*	0.699	**83.2%**	6	0.680	0.472	*79.5%*
Gemini 3.1 Pro (DeepMind, 2026b)	5	0.808	0.582	68.9%	7	0.779	0.621	71.8%	5	0.683	0.476	75.3%
GPT-5.6 Terra	6	0.771	0.595	78.5%	4	0.837	0.699	*80.4%*	10	0.499	0.345	76.8%
Claude Sonnet 4.6 (Anthropic, 2026c)	7	0.769	0.595	66.2%	8	0.742	0.582	65.7%	2	*0.805*	*0.667*	68.4%
Qwen3.7-Max (Qwen3.7, 2026)	8	0.777	0.582	64.3%	6	0.829	0.647	58.5%	9	0.534	0.315	48.2%
Gemini 3.1 FL (DeepMind, 2026a)	9	0.742	0.517	52.2%	10	0.678	0.451	34.7%	7	0.648	0.445	49.3%
GPT-5.4 Mini (OpenAI, 2026b)	10	0.701	0.462	48.6%	12	0.569	0.412	35.3%	12	0.331	0.194	37.4%
Claude Haiku 4.5 (Anthropic, 2025)	11	0.654	0.386	40.5%	9	0.682	0.503	38.7%	8	0.641	0.415	43.8%
GPT-5.6 Luna	12	0.592	0.294	20.6%	11	0.608	0.451	33.1%	11	0.385	0.215	18.7%
Qwen3.6-27B ♠ (Qwen3.6-27B, 2026)	13	0.423	0.192	9.9%	13	0.535	0.362	11.3%	13	0.347	0.176	15.6%

Table 4: The RAVEN-Eval Judge Leaderboards. "Conf." denotes the proportion of pairwise judgments classified as confident for each LMM judge. ♠ denotes an open-weight model. **LMMs are ranked according to the sum of their bootstrap-stabilized *SRCC* and *KRCC* values**.

alignment. We evaluate different combinations of K and M. For $M{=}1$ and $M{=}2$, we consider all three possible single-judge and two-judge combinations, respectively. Since our reliability-based anchor selection requires outputs from all three judges, the $M{=}1$ and $M{=}2$ settings use random anchors, whereas $M{=}3$ includes both our strategy and the random baseline. For each setting, we compute the bootstrap-stabilized *SRCC* against the fully connected 20-model ranking in Tab. 1, the human-annotated ranking, and the *AA* leaderboard. For comparison with the human-annotated ranking, anchors are selected using the same 50 human-annotated tasks. Each configuration is evaluated over 10 independent 15/5 split trials, and the average results are reported in Tab. 3.

When inserting 5 new models, the anchor-based strategy reduces the pairwise annotation cost by 31.6% for $K = 1$ and 18.4% for $K = 2$, while producing rankings close (with *SRCC* over 0.8) to those inferred from the fully connected comparison graph. With a fixed number of anchors, increasing M is more likely to improve ranking reliability. Moreover, our reliability-based anchor selection clearly outperforms random selection across different settings. These results provide empirical support for the large-scale incorporation of future models using representative anchors.

RAVEN-Eval Judge Leaderboards Evaluating AIVGMs also provides a testbed for LMM judges, particularly their ability to recognize fine-grained quality differences among AIGVs. We therefore assess a range of recent open-weight and proprietary LMMs using the same inputs, evaluation protocol, and score-estimation procedure as those used in the *RAVEN-Eval AIVGM Leaderboards*. For each judge, we independently derive an AIVGM ranking and measure its agreement with the human reference using the bootstrap-stabilized *SRCC* and *Kendall's Rank Correlation Coefficient* (*KRCC*). The resulting **RAVEN-Eval Judge Leaderboards**, reported in Tab. 4 and visualized in Fig. 4, benchmark fine-grained AIGV preference-judging capability. We observe that larger LMMs generally outperform smaller ones, while stronger judges also tend to produce higher proportions of confident judgments. The Judge Leaderboards further demonstrate the potential of RAVEN-Eval as a **judge-agnostic evaluation harness**. Several mainstream proprietary LMMs achieve competitive performance within the framework, suggesting that its effectiveness is not tied to a single judge model. This flexibility allows users to select LMM judges according to their evaluation needs and computational budgets, supporting adaptable and scalable evaluation across heterogeneous models.

5 Conclusion

We present **RAVEN-Eval**, an automated framework for evaluating AIVGMs through rubric-guided LMM preference annotations. The **RAVEN-Eval Benchmark** comprises 250 T2V and I2V tasks and over $4,500$ AIGVs. Based on this, we establish the **RAVEN-Eval Leaderboards**, which comprehensively record the capability of 20 SOTA AIVGMs and the evaluation capabilities of 13 LMM judges. We further introduce an **anchor-based model insertion strategy** to control the evaluation cost as the model pool continues to expand. Overall, RAVEN-Eval establishes an **efficient** and **scalable** paradigm for automatic AIVGM evaluation.

AI USE STATEMENT

We used LMMs for task and prompt construction, source-video segment mining, rubric generation, task filtering, candidate-video selection, and pairwise video evaluation, as described in the main text and supplementary materials. Generative AI tools also assisted with manuscript editing and translation, literature discovery, and methodological feedback. The authors reviewed AI-assisted text and suggestions and verified the reported experimental procedures and numerical calculations against the implementation and evaluation records. We assessed the automated evaluation framework through the human preference study described in the paper. The authors take full responsibility for the final manuscript, methodology, and reported results.

ETHICS STATEMENT

RAVEN-Eval supports reliable evaluation of video generation models. Source videos are collected under Creative Commons licenses, and task filtering includes checks for potentially harmful or copyright-restricted content. The human evaluation study was conducted in full compliance with the applicable ITU-T recommendations, with detailed procedures provided in the supplementary materials. Generation-model identities are concealed during LMM evaluation to reduce model-name-based bias.

REPRODUCIBILITY STATEMENT

The main text describes the benchmark construction pipeline, rubric-guided pairwise evaluation, capability estimation, and anchor-based model insertion procedure. The reported results can be reproduced by following the parameter settings and experimental protocols detailed in the supplementary materials, including task construction, rubric definitions, judge input configurations, and human evaluation procedures.

REFERENCES

Anthropic. Claude Haiku 4.5 System Card, October 2025. URL https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf.

Anthropic. Claude Opus 4.6 System Card, February 2026a. URL https://www.anthropic.com/claude-opus-4-6-system-card.

Anthropic. Claude Opus 4.8 System Card, May 2026b. URL https://www.anthropic.com/claude-opus-4-8-system-card.

Anthropic. Claude Sonnet 4.6 System Card, February 2026c. URL https://www.anthropic.com/claude-sonnet-4-6-system-card.

Arena. https://arena.ai/leaderboard, 2026. URL https://arena.ai/leaderboard.

Artificial Analysis. https://artificialanalysis.ai/video/leaderboard, 2026.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=9D2QvO1uWj.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025b.

Emma L Barratt and Nick J Davis. Autonomous sensory meridian response (asmr): a flow-like mental state. *PeerJ*, 3:e851, 2015.

Maosong Cao, Kai Chen, Haodong Duan, Yixiao Fang, Zhiwei Fei, Tong Gao, Jiaye Ge, Mo Li, Hongwei Liu, Junnan Liu, Yuan Liu, Chengqi Lyu, Han Lyu, Ningsheng Ma, Zerun Ma, Yu Sun, Zhiyong Wu, Linchen Xiao, Zhuozhi Xiong, Jun Xu, Haochen Ye, Zhaohui Yu, Yike Yuan, Songyang Zhang, Yufeng Zhao, Fengzhe Zhou, Peiheng Zhou, Dongsheng Zhu, Lin Zhu, and Jingming Zhuo. A universal evaluation platform for large language models. *arXiv preprint arXiv:2605.19276*, 2026.

Siyan Chen et al. Seedance 1.5 Pro: A native audio-visual joint generation foundation model. *arXiv preprint arXiv:2512.13507*, 2025. doi: 10.48550/arXiv.2512.13507. URL https://arxiv.org/abs/2512.13507.

Pierre Dangauthier, Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill through time: Revisiting the history of chess. *Advances in neural information processing systems*, 20, 2007.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082.

DeepMind. Gemini 3.1 Flash-Lite Model Card, March 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/.

DeepMind. Gemini 3.1 Pro Model Card, February 2026b. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Mark E Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(3):377–394, 1999.

Grok Imagine Video. https://docs.x.ai/developers/models/grok-imagine-video, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. doi: 10.48550/arXiv.2601.03233. URL https://arxiv.org/abs/2601.03233.

Hailuo2.3. https://www.minimax.io/news/minimax-hailuo-23, 2025.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, Chi Zhang, Li-jia Li, and Yongxin Ni. Video-Bench: Human-aligned video generation benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18858–18868, 2025. doi: 10.1109/CVPR52734.2025.01757.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 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.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. doi: 10.1109/TPAMI.2025.3633890.

Kling 3.0. https://kling.ai/, 2026. URL https://app.klingai.com/global/quickstart/klingai-video-3-model-user-guide.

Klaus Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70, 1970.

Ilia Kulikov, Chenxi Whitehouse, Tianhao Wu, Yixin Nie, Swarnadeep Saha, Eryk Helenowski, Weizhe Yuan, Olga Golovneva, Jack Lanchantin, Yoram Bachrach, et al. Autodata: An agentic data scientist to create high quality synthetic data. *arXiv preprint arXiv:2606.25996*, 2026.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090.

LTX-2.3. https://ltx.io/model/ltx-2-3, 2026. URL https://ltx.io/model/ltx-2-3.

OpenAI. GPT-5.4 Thinking System Card, March 2026a. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

OpenAI. https://developers.openai.com/api/docs/models/gpt-5.4-mini, 2026b.

OpenAI. GPT-5.5 System Card, April 2026c.

OpenAI. GPT-5.6 System Card, July 2026d. URL https://deploymentsafety.openai.com/gpt-5-6.

PixVerse C1. https://pixverse.ai/en/blog/pixverse-introduces-c1-ai-video-model-for-film-production, 2026.

PixVerse V5.5. https://app.pixverse.ai/, 2025.

PixVerse V6. https://pixverse.ai/en/blog/pixverse-launches-v6-advancing-ai-video-generation, 2026.

Qwen3.6-27B. https://huggingface.co/Qwen/Qwen3.6-27B, 2026.

Qwen3.7. https://qwen.ai/blog?id=qwen3.7, 2026.

Seedance Team et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. doi: 10.48550/arXiv.2604.14148. URL https://arxiv.org/abs/2604.14148.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025.

Veo 3.1. https://deepmind.google/models/veo/, 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf.

Video Arena. https://www.videoarena.tv/leaderboard, 2026.

Jiarui Wang, Huiyu Duan, Guangtao Zhai, Juntong Wang, and Xiongkuo Min. AIGV-Assessor: Benchmarking and evaluating the perceptual quality of text-to-video generation with LMM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18869–18880, 2025.

Jiarui Wang, Huiyu Duan, Ziheng Jia, Yu Zhao, Woo Yi Yang, Zicheng Zhang, Zijian Chen, Juntong Wang, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. LOVE: Benchmarking and evaluating text-to-video generation and video-to-text interpretation. *International Conference on Machine Learning*, 2026.

WanVideo. https://wan.video/, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Hanwei Zhu, Haoning Wu, Yixuan Li, Zicheng Zhang, Baoliang Chen, Lingyu Zhu, Yuming Fang, Guangtao Zhai, Weisi Lin, and Shiqi Wang. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629, 2024.