

RAVEN-Eval: Rubric-Guided Automatic Evaluation for AI Video Generation Models Based on LMM Preference Judgement

Ziheng Jia¹, Jiaying Qian¹, Zicheng Zhang², Xiaorong Zhu¹, Lancheng Gao¹, Xionguo Min^{1*}

¹Shanghai Jiao Tong University

²Shanghai Artificial Intelligence Laboratory
jzhws1@sjtu.edu.cn

Abstract

AI video generation has advanced rapidly and 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 judgement**, 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 and trustworthy evaluation of rapidly evolving AIVGMs.

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

*Corresponding author.

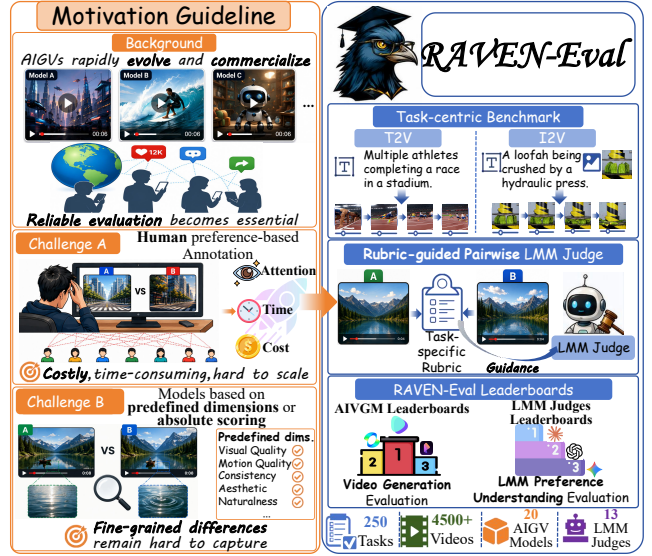


Figure 1: To address the rising cost and limited fine-grained discrimination of existing AIVGM evaluation methods, we introduce **RAVEN-Eval**, a task-centric, rubric-guided LMM pairwise judgment framework for T2V and I2V tasks.

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?

To address this question, we propose **RAVEN-Eval**, a task-centric rubric-guided automatic evaluation framework that employs large multimodal model (LMM) 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:

- 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 the training-free nature and strong **model-harnessing capability** of RAVEN-Eval.

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 automatic 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) 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 score assignment. *VideoPhy* (Bansal et al. 2025a) and *VideoPhy-2* (Bansal et al. 2025b) focus on physical common-sense 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 compelling motivation for our work.

The RAVEN-Eval Benchmark

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 *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 2026b) to generate a **prompt skeleton** defining

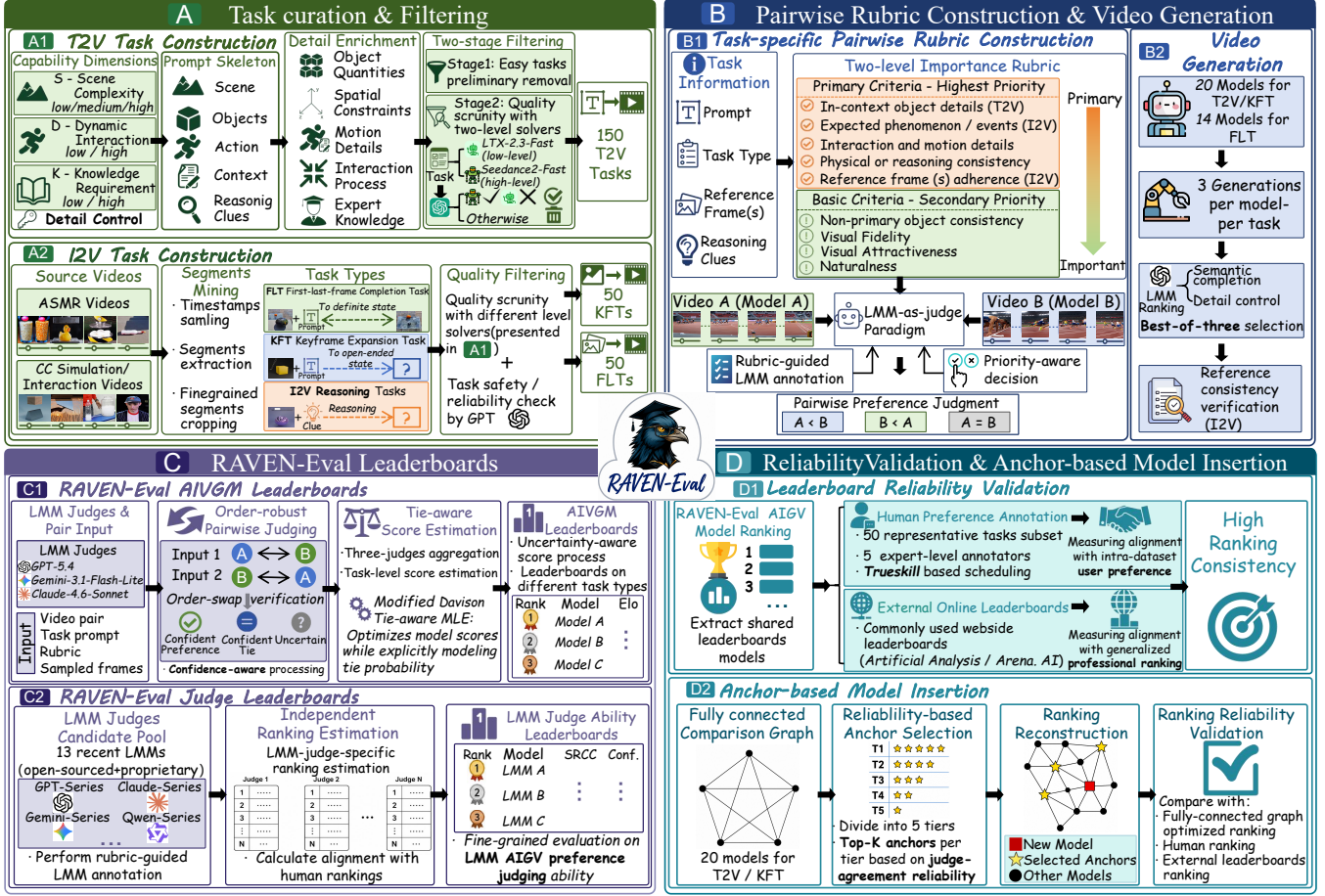


Figure 2: The detailed visualization of the complete RAVEN-Eval construction pipeline.

the core scene and interactions according to its assigned capability dimensions, and then make task **detail enrichment**. 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 task-generation 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 T2V tasks with sufficient discriminative power. 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) and

balancing capability with inference cost, we select *Seedance2 Fast* (Seedance Team et al. 2026) and *LTX 2.3 Fast* (LTX-2.3 2026) as preliminary **solvers** with substantially different capability levels. The former is a high-performance proprietary model, whereas the latter is a smaller and relatively weaker open-weight model. Both solvers generate videos for every task retained after the first-stage filtering. We then use *GPT-5.4-mini* (OpenAI 2026d) 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 3 and *LTX 2.3 Fast* scores at most 2. These tasks exhibit clear performance gaps between solvers with different prior capabilities and are therefore more informative for distinguishing AIVGMs.

I2V Tasks We adopt a separate construction strategy for FLT and KFT. 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 and Davis 2015). Such videos typ-

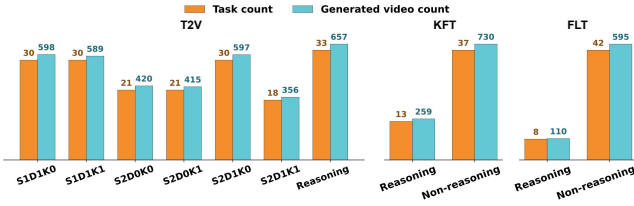


Figure 3: Statistical data of the **RAVEN-Eval Benchmark**. It comprises 250 tasks and 4,669 videos in total.

ically 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 one frame every 3 seconds and provide the sequence to *GPT-5.4-mini* for **segment mining**. The model identifies intervals containing complete physical transformations or object interactions and estimates their temporal boundaries. We then uniformly sample 16 frames from each interval and use the same model to refine the segment boundaries. Segments with insufficient variation between the first and last frames are removed based on pooled pixel-vector similarity. For each retained segment, *GPT-5.4-mini* generates an I2V prompt according to the task type. For **FLTs**, the prompt specifies a coherent transition between the provided first and last frames, with both endpoints strictly aligned to their references and the intermediate process reasonably inferred from the source segment. For **KFTs**, the prompt aligns the initial state with the provided keyframe, while the subsequent event is open-ended, only required to remain logically consistent. We also designate approximately 20% of the I2V tasks as reasoning-based tasks, whose prompts provide only clues derived from the reference images without explicitly stating the expected outcome. These clues must uniquely imply a predictable phenomenon or event, requiring the AIVGM to infer and depict the intended content. Following the Stage-2 T2V filtering protocol, we use *Seedance2 Fast*, *LTX 2.3 Fast*, and *GPT-5.4-mini* to filter the I2V tasks. Finally, a 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 **effective rubrics** to accurately 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 task-specific preference evaluation rubric, consisting of primary basic criteria. The primary compliance criteria mainly emphasize **how accurately the generated video reproduces the detailed**

requirements specified in the prompt. 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.

RAVEN-Eval Leaderboards

RAVEN-Eval AIVGM Leaderboards After constructing the benchmark, we perform LMM-as-a-judge pairwise 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 intra-dataset selection bias, these judges are chosen **a priori** based on the trade-off between their recognized multimodal understanding capabilities and inference costs. The judgement input content includes 8 uniformly sampled frames from each video, the prompt, the task-specific rubric, and, for I2V tasks, the corresponding visual reference. Detailed settings are provided in *Supp. Sec. A.5*.

Because some AIGV pairs remain difficult to distinguish even 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):

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

Here, $\mathcal{P}$ denotes the set of valid pairwise comparisons for a given task, where $p = (i, j)$ represents a comparison between two AIGV models. N and M denote the numbers of AIGVMs to be evaluated 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 ($\text{Elo}_{i,c} = 1000 + \frac{400}{\ln 10} \bar{s}_{i,c}$) for intuitive presentation in Tab. 1. The process of building the leaderboard is also shown in Fig. 2.

Reliability Verification To verify the reliability of the *RAVEN-Eval AIGVM Leaderboards*, we conduct human pref-

erence experiments on 50 tasks, including 30 T2V tasks, 10 KFTs, and 10 FLT. T2V tasks are sampled to preserve the original distribution of the S , D , and K levels, while I2V 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, Minka, and Graepel 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 AIGVM 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 (detailed in *Supp.*) and compute the corresponding *SRCC*. The overall validation process is illustrated in Fig. 2, 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 AIGVMs, averaged across dimensions for each video, and then averaged across videos to obtain the final score of each AIGVM. The

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

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 (SRCC) on the effectiveness of task-specific rubrics. The term “H.” is short for “Human”.

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.

Anchor-Based New Model Insertion

Although LMM evaluation substantially reduces costs, the number of pairwise comparisons still grows quadratically with the size of the model pool. Maintaining a fully connected comparison graph is therefore impractical for continuously updating the model pool. Inspired by (Zhu et al. 2024), we introduce an anchor-based insertion strategy (also shown in Fig. 2), in which newly added models are compared only with a subset of anchors from the existing model pool.

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 a initial pool and treat the remaining 5 as newly inserted models. For each candidate

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 (SRCC) under different K, M settings using either the proposed anchor-selection or random selection. “Full” uses the ranking from the complete comparison graph as the reference. “Rand.” denotes random pick.

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.

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

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.

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 alignment.

We evaluate different combinations of K , M , and anchor-selection strategy. For $M=1$ and $M=2$, we consider all three possible single-judge and two-judge combinations, respectively, and report their averaged results. 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 *SRCC* against the fully connected 20-model ranking in Tab. 1, the human-annotated ranking, and the external 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 34.2% for $K=1$ and 21.1% 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 *SRCC* and *Kendall’s Rank Correlation Coefficient (KRCC)*. The resulting **RAVEN-Eval Judge Leaderboards**, reported in Tab. 4 and visualized in Fig. 2, benchmark fine-grained AIGV preference-judging capability. We observe that larger-size 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 for effective **LMM judge model harnessing**. We observe that several mainstream proprietary LMMs achieve competitive performance, indicating that the framework is not overly sensitive to a particular judge choice. This flexibility provides a foundation for harnessing heterogeneous LMMs according to user-specific requirements and for further building more adaptable and scalable evaluation systems.

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 intro-

duce 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.

References

- Anthropic. 2025. Claude Haiku 4.5 System Card.
- Anthropic. 2026a. Claude Opus 4.6 System Card.
- Anthropic. 2026b. Claude Opus 4.8 System Card.
- Anthropic. 2026c. Claude Sonnet 4.6 System Card.
- Arena. 2026. <https://arena.ai/leaderboard>.
- Artificial Analysis. 2026. <https://artificialanalysis.ai/video/leaderboard>.
- Bansal, H.; Lin, Z.; Xie, T.; Zong, Z.; Yarom, M.; Bitton, Y.; Jiang, C.; Sun, Y.; Chang, K.-W.; and Grover, A. 2025a. Evaluating Physical Commonsense for Video Generation. In *International Conference on Learning Representations*.
- Bansal, H.; Peng, C.; Bitton, Y.; Goldenberg, R.; Grover, A.; and Chang, K.-W. 2025b. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. *arXiv preprint arXiv:2503.06800*.
- Barratt, E. L.; and Davis, N. J. 2015. Autonomous Sensory Meridian Response (ASMR): a flow-like mental state. *PeerJ*, 3: e851.
- Cao, M.; Chen, K.; Duan, H.; Fang, Y.; Fei, Z.; Gao, T.; Ge, J.; Li, M.; Liu, H.; Liu, J.; Liu, Y.; Lyu, C.; Lyu, H.; Ma, N.; Ma, Z.; Sun, Y.; Wu, Z.; Xiao, L.; Xiong, Z.; Xu, J.; Ye, H.; Yu, Z.; Yuan, Y.; Zhang, S.; Zhao, Y.; Zhou, F.; Zhou, P.; Zhu, D.; Zhu, L.; and Zhuo, J. 2026. A Universal Evaluation Platform for Large Language Models. *arXiv preprint arXiv:2605.19276*.
- Chen, S.; et al. 2025. Seedance 1.5 Pro: A Native Audio-Visual Joint Generation Foundation Model. *arXiv preprint arXiv:2512.13507*.
- Dangauthier, P.; Herbrich, R.; Minka, T.; and Graepel, T. 2007. Trueskill through time: Revisiting the history of chess. *Advances in neural information processing systems*, 20.
- Davidson, R. R. 1970. On Extending the Bradley-Terry Model to Accommodate Ties in Paired Comparison Experiments. *Journal of the American Statistical Association*, 65(329): 317–328.
- DeepMind. 2026a. Gemini 3.1 Flash-Lite Model Card.
- DeepMind. 2026b. Gemini 3.1 Pro Model Card.
- Glickman, M. E. 1999. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 48(3): 377–394.
- Grok Imagine Video. 2026. <https://docs.x.ai/developers/models/grok-imagine-video>.
- HaCohen, Y.; Brazowski, B.; Chiprut, N.; Bitterman, Y.; Kvochko, A.; Berkowitz, A.; Shalem, D.; Lifschitz, D.; Moshe, D.; Porat, E.; Richardson, E.; et al. 2026. LTX-2: Efficient Joint Audio-Visual Foundation Model. *arXiv preprint arXiv:2601.03233*.
- Hailuo2.3. 2025. <https://www.minimax.io/news/minimax-hailuo-23>.
- Han, H.; Li, S.; Chen, J.; Yuan, Y.; Wu, Y.; Deng, Y.; Leong, C. T.; Du, H.; Fu, J.; Li, Y.; Zhang, J.; Zhang, C.; Li, L.-j.; and Ni, Y. 2025. Video-Bench: Human-Aligned Video Generation Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18858–18868.
- 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, B. Y.; and Chen, W. 2024. VideoScore: Building Automatic Metrics to Simulate Fine-Grained Human Feedback for Video Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2105–2123. Association for Computational Linguistics.
- Herbrich, R.; Minka, T.; and Graepel, T. 2006. TrueSkill: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, volume 19.
- 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.; and Liu, Z. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.
- Huang, Z.; Zhang, F.; Xu, X.; He, Y.; Yu, J.; Dong, Z.; Ma, Q.; Chanpaisit, N.; Si, C.; Jiang, Y.; Wang, Y.; Chen, X.; Chen, Y.-C.; Wang, L.; Lin, D.; Qiao, Y.; and Liu, Z. 2025. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Kling 3.0. 2026. <https://kling.ai/>.
- Krippendorff, K. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1): 61–70.
- Kulikov, I.; Whitehouse, C.; Wu, T.; Nie, Y.; Saha, S.; Helenowski, E.; Yuan, W.; Golovneva, O.; Lanchantin, J.; Bachrach, Y.; et al. 2026. Autodata: An agentic data scientist to create high quality synthetic data. *arXiv preprint arXiv:2606.25996*.
- Liu, Y.; Cun, X.; Liu, X.; Wang, X.; Zhang, Y.; Chen, H.; Liu, Y.; Zeng, T.; Chan, R.; and Shan, Y. 2024. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22139–22149.
- LTX-2.3. 2026. <https://ltx.io/model/ltx-2-3>.
- OpenAI. 2026a. GPT-5.4 Thinking System Card.
- OpenAI. 2026b. GPT-5.5 System Card. <https://openai.com/index/gpt-5-5-system-card/>.
- OpenAI. 2026c. GPT-5.6 System Card.
- OpenAI. 2026d. <https://developers.openai.com/api/docs/models/gpt-5.4-mini>.
- PixVerse C1. 2026. <https://pixverse.ai/en/blog/pixverse-introduces-c1-ai-video-model-for-film-production>.
- PixVerse V5.5. 2025. <https://app.pixverse.ai/>.

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

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

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

Seedance Team; et al. 2026. Seedance 2.0: Advancing Video Generation for World Complexity. *arXiv preprint arXiv:2604.14148*.

Sun, K.; Huang, K.; Liu, X.; Wu, Y.; Xu, Z.; Li, Z.; and Liu, X. 2025. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8406–8416.

Veo 3.1. 2025. <https://deepmind.google/models/veo/>.

Video Arena. 2026. <https://www.videoarena.tv/leaderboard>.

Wang, J.; Duan, H.; Jia, Z.; Zhao, Y.; Yang, W. Y.; Zhang, Z.; Chen, Z.; Wang, J.; Xing, Y.; Zhai, G.; and Min, X. 2026. LOVE: Benchmarking and Evaluating Text-to-Video Generation and Video-to-Text Interpretation. *International Conference on Machine Learning*.

Wang, J.; Duan, H.; Zhai, G.; Wang, J.; and Min, X. 2025. 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*, 18869–18880.

WanVideo. 2026. <https://wan.video/>.

Zheng, D.; Huang, Z.; Liu, H.; Zou, K.; He, Y.; Zhang, F.; Gu, L.; Zhang, Y.; He, J.; Zheng, W.-S.; Qiao, Y.; and Liu, Z. 2025. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*.

Zhu, H.; Wu, H.; Li, Y.; Zhang, Z.; Chen, B.; Zhu, L.; Fang, Y.; Zhai, G.; Lin, W.; and Wang, S. 2024. Adaptive image quality assessment via teaching large multimodal model to compare. *Advances in Neural Information Processing Systems*, 37: 32611–32629.

A Supplementary Materials

A.1 Information on the Evaluated AIVGMs

This subsection provides supplementary information on the AIVGMs evaluated in RAVEN-Eval, including their open-source status, year of first public release, and official model or product pages. A model is regarded as open source when its official model weights are publicly available for download, even if its complete training code or training data are unavailable.

1. **Seedance 2.** Seedance 2 is a closed-source model first publicly released in 2026. **Official Link:** <https://seed.bytedance.com/en/seedance2.0>.
2. **Kling 3.0 Pro.** Kling 3.0 Pro is a closed-source model first publicly released in 2026. **Official Link:** <https://home.klingai.com/>.
3. **Grok-Image-Video.** Grok-Image-Video, officially provided as Grok Imagine Video, is a closed-source model first publicly released in 2026. **Official Link:** <https://docs.x.ai/developers/models/grok-image-video>.
4. **HappyHorse V1.0.** HappyHorse V1.0 is a closed-source model first publicly released in 2026. **Official Link:** <https://www.alibabacloud.com/blog/603068>.
5. **Kling 3.0 Omni.** Kling 3.0 Omni is a closed-source model first publicly released in 2026. **Official Link:** <https://home.klingai.com/>.
6. **Seedance 2 Fast.** Seedance 2 Fast is a closed-source model first publicly released in 2026. **Official Link:** <https://www.volcengine.com/product/dreamart>.
7. **PixVerse C1.** PixVerse C1 is a closed-source model first publicly released in 2026. **Official Link:** <https://pixverse.ai/en/blog/pixverse-introduces-c1-ai-video-model-for-film-production>.
8. **Kling 3.0 Std.** Kling 3.0 Std is a closed-source model first publicly released in 2026. **Official Link:** <https://home.klingai.com/>.
9. **WAN 2.7.** WAN 2.7 is a closed-source model first publicly released in 2026. **Official Link:** <https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model>.
10. **PixVerse V6.** PixVerse V6 is a closed-source model first publicly released in 2026. **Official Link:** <https://pixverse.ai/en/blog/pixverse-launches-v6-advancing-ai-video-generation>.
11. **Veo 3.1 Fast.** Veo 3.1 Fast is a closed-source model first publicly released in 2025. **Official Link:** <https://blog.google/innovation-and-ai/products/veo-updates-flow/>.
12. **Veo 3.1 Lite.** Veo 3.1 Lite is a closed-source model first publicly released in 2026. **Official Link:** <https://cloud.google.com/blog/products/ai-machine-learning/veo-3-1-lite-and-a-new-veo-upscaling-capability-on-vertex-ai>.
13. **LTX 2.3 Pro.** LTX 2.3 Pro was first publicly released in 2026 and is regarded as open source because its official model weights are publicly available. **Official Link:** <https://ltx.io/blog/ltx-2-3-release>.

14. **WAN 2.6.** WAN 2.6 is a closed-source model first publicly released in 2025. **Official Link:** <https://www.alibabacloud.com/help/en/model-studio/video-generate-edit-model>.
15. **PixVerse V5.5.** PixVerse V5.5 is a closed-source model first publicly released in 2025. **Official Link:** <https://docs.platform.pixverse.ai/changelogs-906383m0>.
16. **LTX 2.3 Fast.** LTX 2.3 Fast was first publicly released in 2026 and is regarded as open source because its official model weights are publicly available. **Official Link:** <https://ltx.io/blog/ltx-2-3-release>.
17. **Seedance 1.5 Pro.** Seedance 1.5 Pro is a closed-source model first publicly released in 2025. **Official Link:** <https://seed.bytedance.com/en/seedance1.5>.
18. **Hailuo 2.3.** Hailuo 2.3 is a closed-source model first publicly released in 2025. **Official Link:** <https://www.minimax.io/news/minimax-hailuo-23>.
19. **LTX 2.0 Pro.** LTX 2.0 Pro was first publicly released in 2025 and is regarded as open source because its official model weights are publicly available. **Official Link:** <https://ltx.io/newsroom/ltx-2-foundation-ai-video-model-is-released>.
20. **LTX 2.0 Fast.** LTX 2.0 Fast was first publicly released in 2025 and is regarded as open source because its official model weights are publicly available. **Official Link:** <https://ltx.io/newsroom/ltx-2-foundation-ai-video-model-is-released>.

A.2 Information on the Evaluated LMMs

This subsection provides supplementary information on the LMM judges evaluated in RAVEN-Eval, including their open-source status, year of first public release, and official model or product pages. Following the criterion used for the evaluated AIVGMs, a model is regarded as open source when its official model weights are publicly available for download, even if its complete training code or training data are unavailable.

1. **Claude Opus 4.6.** Claude Opus 4.6 is a closed-source model first publicly released in 2026. **Official Link:** <https://www.anthropic.com/news/claude-opus-4-6>.
2. **GPT-5.4.** GPT-5.4 is a closed-source model first publicly released in 2026. **Official Link:** <https://openai.com/index/introducing-gpt-5-4/>.
3. **GPT-5.6 Sol.** GPT-5.6 Sol is a closed-source model first publicly released in 2026 as the flagship member of the GPT-5.6 family. **Official Link:** <https://openai.com/index/gpt-5-6/>.
4. **Claude Opus 4.8.** Claude Opus 4.8 is a closed-source model first publicly released in 2026. **Official Link:** <https://www.anthropic.com/news/claude-opus-4-8>.
5. **Gemini 3.1 Pro.** Gemini 3.1 Pro is a closed-source model first publicly released in 2026. **Official Link:** <https://deepmind.google/models/model-cards/gemini-3-1-pro/>.
6. **GPT-5.6 Terra.** GPT-5.6 Terra is a closed-source model first publicly released in 2026 as the balanced, lower-cost member of the GPT-5.6 family. **Official Link:** <https://openai.com/index/gpt-5-6/>.

7. **Claude Sonnet 4.6.** Claude Sonnet 4.6 is a closed-source model first publicly released in 2026. **Official Link:** <https://www.anthropic.com/news/claude-sonnet-4-6>.
8. **Qwen3.7-Max.** Qwen3.7-Max is a closed-source model first publicly released in 2026, with access provided through Qwen’s hosted services rather than downloadable model weights. **Official Link:** <https://qwen.ai/blog?id=qwen3.7>.
9. **Gemini 3.1 Flash-Lite.** Gemini 3.1 Flash-Lite is a closed-source model first publicly released in 2026. **Official Link:** <https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/>.
10. **GPT-5.4 Mini.** GPT-5.4 Mini is a closed-source model first publicly released in 2026 as a smaller and more efficient member of the GPT-5.4 family. **Official Link:** <https://openai.com/index/introducing-gpt-5-4-mini-and-nano/>.
11. **Claude Haiku 4.5.** Claude Haiku 4.5 is a closed-source model first publicly released in 2025. **Official Link:** <https://www.anthropic.com/news/claude-haiku-4-5>.
12. **GPT-5.6 Luna.** GPT-5.6 Luna is a closed-source model first publicly released in 2026 as the fastest and most cost-efficient member of the GPT-5.6 family. **Official Link:** <https://openai.com/index/gpt-5-6/>.
13. **Qwen3.6-27B.** Qwen3.6-27B was first publicly released in 2026 and is regarded as open source because its official model weights are publicly available under the Apache 2.0 license. **Official Link:** <https://huggingface.co/Qwen/Qwen3.6-27B>.

A.3 Prompt Summary

The Basic Criteria in Task-Specific Rubrics

1. Static non-target objects and background elements should remain consistent in both shape and quantity throughout the video.
2. Unless otherwise specified by the task instruction, the video should preserve practical realism as much as possible. It should follow real-world common sense, especially ensuring that any text appearing in the video is realistic, legible, and understandable. It should also follow real-world physical, material, and reaction laws. In particular, for videos involving material reactions or object interactions, dynamic objects undergoing reactions or interactions should exhibit logically coherent transformation processes. The total amount of material should be conserved as strictly as possible, without unrealistic material generation or abrupt changes.
3. For videos involving dynamic processes, the action or reaction process should be complete and reasonable. Unless otherwise specified by the task instruction, all actions or reactions should appear natural and consistent with real-world logic.
4. After satisfying the above three criteria, which have higher priority, the generated video should have as much aesthetic value and visual appeal as possible.

5. If a primary criteria instruction conflicts with the above four criteria, such as a scene that explicitly does not require realistic physical behavior, the task-specific instruction should take priority. The above four criteria should then be treated as secondary evaluation standards.

T2V LMM Judge Prompt

T2V task: You are a professional AI text-to-video quality evaluation expert. You are asked to perform preference comparison for a series of AIGC video triples generated by different text-to-video models according to the given evaluation criteria. Each triple includes two AIGC videos generated by different models, the T2V prompt input for this task, and the specific comparison criteria. The criteria contain two priority levels: primary considerations and basic criteria (lower priority than the primary considerations).

When judging preference, you should first strictly compare according to the primary considerations. If a preference can be directly determined, use that preference as the output result. If the primary considerations are insufficient, jointly judge according to the basic criteria.

The final output format is label only: better / worse / similar, indicating whether the first video is better than, worse than, or indistinguishable from the second video under the evaluation criteria.

When determining the preference, you must: 1. Strictly base the evaluation result and related inference on the evaluation criteria and their priority order. Do not infer from knowledge outside these criteria. 2. Strictly base the evaluation result on the video pair content and the prompt content. Do not use any other modality or information. 3. Do not consider audio. Audio-related content is unnecessary and prohibited. Only visual content, the prompt, and the given evaluation criteria should be considered.

This comparison only involves the following two generated videos: - First video: generated_video_a, model {MODEL_A} - Second video: generated_video_b, model {MODEL_B}

prompt: {PROMPT}

RUBRIC: {RUBRIC}

Strictly follow the guideline above. Do not output reasoning, drafts, analysis, step-by-step explanations, self-correction, or statements about how you will answer. This run uses the label_only output format. The final response must be exactly one of the following three labels: “better”, “worse”, or “similar”. Do not output any reason or extra text.

KFT LMM Judge Prompt

Keyframe extension task: You are a professional AI-generated video quality evaluation expert. You are asked to perform preference comparison for a series of AIGC video triples generated by different models according to the given evaluation criteria. Each triple

includes two AIGC videos generated by different models and the reference first frame used for the keyframe extension task (the first frame sampled from the reference video is used as the input for the keyframe extension task). The task prompt input and the specific comparison criteria are also provided. The criteria contain two priority levels: primary considerations and basic criteria.

When judging preference, you should first strictly compare according to the primary considerations. If a preference can be directly determined, use that preference as the output result. If the primary considerations are insufficient, jointly judge according to the basic criteria.

The final output format is label only: better / worse / similar, indicating whether the first video is better than, worse than, or indistinguishable from the second video under the evaluation criteria.

When determining the preference, you must: 1. Strictly base the evaluation result and related inference on the evaluation criteria and their priority order. Do not infer from knowledge outside these criteria. 2. Strictly base the evaluation result on the video pair content and the prompt content. Do not use any other modality or information. 3. Do not consider audio. Audio-related content is unnecessary and prohibited. Only visual content, the prompt, and the given evaluation criteria should be considered.

This comparison only involves the following two generated videos: - First video: generated_video_a, model {MODEL_A} - Second video: generated_video_b, model {MODEL_B}

prompt: {PROMPT}

RUBRIC: {RUBRIC}

Strictly follow the guideline above. Do not output reasoning, drafts, analysis, step-by-step explanations, self-correction, or statements about how you will answer. This run uses the label_only output format. The final response must be exactly one of the following three labels: “better”, “worse”, or “similar”. Do not output any reason or extra text.

FLT LMM Judge Prompt

First-last-frame completion task: You are a professional AI-generated video quality evaluation expert. You are asked to perform preference comparison for a series of AIGC video triples generated by different models according to the given evaluation criteria. Each triple includes two AIGC videos generated by different models and the reference first and last frames used for the first-last-frame completion task (the first and last frames sampled from the reference video are used as the input for the first-last-frame completion task). The task prompt input and the specific comparison criteria are also provided. The criteria contain two priority levels: primary considerations and basic criteria.

When judging preference, you should first strictly compare according to the primary considerations. If a pref-

erence can be directly determined, use that preference as the output result. If the primary considerations are insufficient, jointly judge according to the basic criteria.

The final output format is label only: better / worse / similar, indicating whether the first video is better than, worse than, or indistinguishable from the second video under the evaluation criteria.

When determining the preference, you must: 1. Strictly base the evaluation result and related inference on the evaluation criteria and their priority order. Do not infer from knowledge outside these criteria. 2. Strictly base the evaluation result on the video pair content and the prompt content. Do not use any other modality or information. 3. Do not consider audio. Audio-related content is unnecessary and prohibited. Only visual content, the prompt, and the given evaluation criteria should be considered.

This comparison only involves the following two generated videos: - First video: generated_video_a, model {MODEL_A} - Second video: generated_video_b, model {MODEL_B}

prompt: {PROMPT}

RUBRIC: {RUBRIC}

Strictly follow the guideline above. Do not output reasoning, drafts, analysis, step-by-step explanations, self-correction, or statements about how you will answer. This run uses the label_only output format. The final response must be exactly one of the following three labels: “better”, “worse”, or “similar”. Do not output any reason or extra text.

Forward and Reversed Evaluation In forward evaluation, generated_video_a is used as the first video, and generated_video_b is used as the second video.

In reversed evaluation, the input order of the two videos is swapped: the original generated_video_b is used as the first video, and the original generated_video_a is used as the second video.

For both forward and reversed evaluation, the output label is always relative to the current input order: “better” means the first video is better, “worse” means the first video is worse, and “similar” means there is no clear quality difference between the two videos.

Prompt for GPT-5.4-mini in the task quality filtering stage

You are an expert evaluator for text-to-video generation tasks. Your goal is to assess how well a candidate video fulfills the given task on a five-point ordinal scale.

You must evaluate the video independently. Do not compare it with videos generated by other models, and do not infer the identity or expected capability of the generation model.

Input

Task type: T2V TASK TYPE

Generation prompt: GENERATION PROMPT

Specified task details: DETAIL REQUIREMENTS

Hidden expected outcome: HIDDEN EXPECTED OUTCOME

Candidate video: VIDEO INPUT

The hidden expected outcome is provided only for reasoning-based tasks. If it is marked as "None", evaluate the video solely according to the generation prompt and specified task details.

Evaluation Objective

Assess prompt fulfillment by jointly considering:

1. Core semantic completion - Whether the principal subjects, objects, scene, and event described in the prompt are present. - Whether the main action, interaction, transformation, or intended outcome is correctly realized. - Whether the generated content preserves the essential meaning of the task rather than merely depicting a related scene.

2. Realization of specified details - Whether explicitly required object quantities, attributes, spatial relations, and compositional constraints are satisfied. - Whether required actions, motion trajectories, interaction sequences, and state transitions are visibly realized. - Whether domain-specific objects, operations, physical phenomena, material responses, or professional details are depicted correctly. - For reasoning-based tasks, whether the video presents the uniquely expected event or outcome implied by the clues.

Core semantic completion has higher priority than secondary details. A visually attractive video that fails to realize the principal task should receive a low score. Conversely, minor visual defects should not dominate the judgment when the core task and most specified details are correctly completed.

Only evaluate content that is visibly supported by the video. Do not assume that an unobserved action or event occurred outside the displayed duration. For dynamic tasks, showing only the initial or final state is insufficient when the prompt explicitly requires an interaction or transformation process.

Scoring Scale

Score 5 – Excellent fulfillment The video correctly realizes the complete core task and nearly all specified details. The main subjects, interactions, spatial relations, and expected phenomena are clearly and coherently presented. Only negligible imperfections are present.

Score 4 – Strong fulfillment The core task is correctly completed, and most important details are realized. Minor omissions, inaccuracies, or local inconsistencies are present, but they do not substantially affect the intended scene or event.

Score 3 – Adequate fulfillment The principal semantic content and main task objective are recognizable and substantially completed. However, several specified details are missing or inaccurate, or the process contains noticeable defects. The video remains a valid realization of the task.

Score 2 – Weak fulfillment The video only partially realizes the task. The main event, interaction, or expected phenomenon is incomplete, incorrect, or insufficiently visible, and multiple important details are absent or contradicted. The video may depict a related scene but does not adequately complete the requested task.

Score 1 – Failed fulfillment The video is unrelated to the prompt, omits the principal subjects or event, seriously contradicts the task, or is too corrupted or incomplete to evaluate as a valid realization.

Decision Rules

- If the principal task or expected event is absent or fundamentally incorrect, assign a score no higher than 2. - If the core task is completed but several secondary details are missing, distinguish between Scores 3 and 4 according to the importance and number of the missing details. - Do not reduce the score solely because of general visual-quality defects unless they prevent recognition of the required content or violate an explicit task requirement. - Do not reward additional content that is not requested if it does not improve fulfillment of the specified task. - Use an integer score only.

Output Format

Return only the following JSON object:

"score": 1, "core semantic completion": "complete — partial — failed", "detail realization": "high — moderate — low", "missing or incorrect requirements": "Briefly list each important missing or incorrect requirement", "rationale": "Provide a concise explanation grounded only in visible evidence from the video."

Prompt for T2V Prompt-Skeleton Construction

You are an expert task designer for evaluating advanced AI video generation models. Given a specified scene-complexity level S , dynamic-interaction level D , domain-knowledge level K , and reasoning-task indicator R , your task is to construct a text-to-video prompt skeleton with a clear visual objective and sufficient potential to distinguish models with different capabilities. At this stage, the prompt skeleton should specify only the core scene, principal subjects and objects, primary state or interaction, and necessary contextual information. Do not introduce excessive spatial, motion, procedural, or domain-specific details, as these requirements will be added during the subsequent detail-enrichment stage.

Input Variables. The scene-complexity level S controls the number of subjects or objects and the complexity of their spatial organization. $S0$ denotes a simple composition containing a single subject or only a few objects with straightforward spatial relationships. $S1$ denotes a scene containing multiple distinguishable subjects or objects with nontrivial spatial relationships, role assignments, or compositional requirements. $S2$ denotes a scene containing numerous subjects or objects organized through complex spatial structures, group relationships, or multilayer compositions.

The dynamic-interaction level D controls the complexity of motion, interaction, and temporal evolution. $D0$ denotes a static or nearly static scene that may contain only limited and simple motion, without a complex interaction process. $D1$ denotes a clearly observable dynamic process involving object interactions, coordinated actions, or multistage

state transitions, for which temporal continuity and causal consistency are essential.

The domain-knowledge level K controls whether correct generation requires specialized knowledge. $K0$ denotes an everyday or general-purpose scenario that does not depend on specialized disciplinary knowledge, professional operations, or material mechanisms. $K1$ denotes a task whose correct visual realization requires domain-specific knowledge, such as physical phenomena, chemical reactions, material transformations, professional sports movements, experimental procedures, text or design conventions, specialized photography, or stage performance.

The reasoning-task indicator R specifies whether the expected visual event is stated explicitly. $R0$ denotes a non-reasoning task in which the required action, process, phenomenon, or final state may be described directly. $R1$ denotes a reasoning-based task in which the public generation prompt may contain only clues from which the expected event or outcome can be uniquely inferred using common-sense or relevant domain knowledge. For an $R1$ task, the expected outcome must not be directly disclosed in the public prompt.

Task Profiles for Different S - D - K Combinations. For $S0$ - $D0$ - $K0$, construct an everyday static scene with a simple subject configuration and uncomplicated composition. For $S1$ - $D0$ - $K0$, construct an everyday static scene containing multiple subjects or objects, with emphasis on basic composition and interpretable object relationships. For $S2$ - $D0$ - $K0$, construct a complex everyday static scene containing many subjects or objects, emphasizing compositional plausibility, spatial naturalness, and subject consistency while imposing minimal professional-detail requirements.

For $S0$ - $D1$ - $K0$, construct an everyday dynamic process involving only a small number of subjects or objects and a simple composition. For $S1$ - $D1$ - $K0$, construct an everyday dynamic interaction involving multiple subjects or objects, with emphasis on action coordination and plausible object interactions. For $S2$ - $D1$ - $K0$, construct a complex and highly dynamic everyday scene involving multiple subjects, coordinated group motion, or concurrent object interactions, while avoiding unnecessary professional knowledge requirements.

For $S0$ - $D0$ - $K1$, construct a professional static scene containing a single subject or only a few objects. Representative scenarios include standardized text presentation, professional product display, a specialized model pose, or a single-subject composition governed by explicit aesthetic conventions. The task should emphasize professional depiction of the principal subject, overall visual style, and compositional quality.

For $S1$ - $D0$ - $K1$, construct a professional static scene containing several subjects or objects. Representative scenarios include arrangements of multiple products, multi-object typography, posed groups of performers, or combinations of professional instruments. The task should emphasize limited static relationships among the principal objects, coherent composition, and accurate professional depiction.

For $S2$ - $D0$ - $K1$, construct a complex professional static scene containing numerous subjects or objects. Representative scenarios include densely arranged professional products, group poses, sculpture ensembles, or laboratory platforms containing multiple instruments. The task should emphasize object-level consistency, structurally plausible arrangement, coherent group organization, and accurate pro-

fessional details.

For $S0$ - $D1$ - $K1$, construct a professional dynamic scene involving a single principal object or a small number of objects in a simple composition. Representative scenarios include a hydraulic press deforming a material, a red-hot metal ball penetrating a substance, a glass bottle rolling down stairs, soap being squeezed or cut, or text being written dynamically. The task should emphasize continuity of the complete interaction process, plausible presentation of the professional setting, and completeness of the required action.

For $S1$ - $D1$ - $K1$, construct a moderately complex professional dynamic scene involving multiple subjects or objects. Representative scenarios include chemical reactions, mechanics experiments, air-pressure demonstrations, optical experiments, or multi-object material interactions. For scientific tasks, the task definition must specify the relevant substances, apparatus, experiment, or mechanism and the expected observable phenomenon with sufficient precision.

For $S2$ - $D1$ - $K1$, construct the most complex category of professional and highly dynamic tasks. Representative scenarios include professionally filmed sprint races, coordinated multi-person dance performances, aerial formation maneuvers, advanced group techniques such as leaf-like descending formations, or complex multi-object mechanics experiments. The task must specify the relevant professional action, experiment, or phenomenon sufficiently clearly to support objective evaluation.

Task-Design Requirements The task must be realizable within the short durations commonly supported by mainstream AI video generation models and should present a complete and observable state, interaction, or event within that duration. It must provide clear visual evidence for evaluation and must not primarily depend on audio, dialogue, explanatory subtitles, or information unavailable within the video. The task should meaningfully distinguish advanced models in terms of subject consistency, composition, motion, interaction, physical plausibility, material behavior, or professional knowledge. Avoid tasks that rely mainly on abstract concepts, subjective emotions, or outcomes that cannot be verified visually. Avoid real-person identity replication, copyrighted characters, brand-specific imitation, unsafe procedural instructions, or other content unsuitable for a public benchmark.

A $K1$ task must contain genuine professional requirements that materially affect the correctness of the generated video. Merely placing an otherwise ordinary event in a laboratory, stadium, studio, or other professional-looking environment is insufficient to satisfy $K1$. An $R1$ task must provide clues that uniquely determine a visually observable outcome. If two or more substantially different outcomes remain equally plausible, the task is invalid and must be redesigned.

Output Format. Return the result strictly using the following fields.

Task Level: Specify the assigned values of S , D , K , and R .

Core Scene: Describe the environment and fundamental situation of the task in one concise sentence.

Subjects and Objects: Identify the principal subjects, key objects, and their basic roles in the scene.

Core State or Interaction: Describe the principal static state, action, transformation, or interaction that the video must present.

Scene Context: Provide only the environmental, temporal, locational, or situational context required to understand the

task. Do not add irrelevant decorative details.

Reasoning Clues: Complete this field only when $R1$ is specified. State the clues that may appear in the public generation prompt without revealing the expected outcome. For an $R0$ task, output “None.”

Hidden Expected Outcome: Describe the event, phenomenon, process, or final state that constitutes the correct realization of the task. This field is intended only for subsequent task filtering and rubric construction. It must not be copied directly into the public prompt for an $R1$ task.

Core Evaluation Focus: State two to four principal capabilities assessed by the task, such as compositional consistency, motion continuity, interaction accuracy, physical plausibility, material behavior, or professional-detail realization.

Prompt Skeleton: Integrate the core scene, principal subjects and objects, core state or interaction, and necessary context into one concise and natural text-to-video generation prompt. Do not include metadata or terms such as S , D , K , R , evaluation, benchmark, rubric, or model capability. For an $R1$ task, include only the permitted reasoning clues and do not explicitly reveal the hidden expected outcome.

Before returning the result, verify that the generated task matches the specified S , D , K , and R levels; that its key requirements can be evaluated from visible video evidence; that it can be completed within a short video; and that any reasoning-based outcome is uniquely inferable without being explicitly disclosed.

Prompt for T2V Prompt Detail Enrichment

You are an expert in refining prompts for the evaluation of advanced AI video generation models. Your task is to enrich a given prompt skeleton according to its scene-complexity level S , dynamic-interaction level D , domain-knowledge level K , and reasoning-task indicator R , without altering its core semantics. The resulting prompt should contain explicit, observable, and objectively assessable requirements while preserving the original scene, principal subjects, and intended event. You must also produce structured task records that can subsequently be used for quality filtering and rubric construction.

The input task has already passed preliminary filtering and belongs to one of the following combinations: $S1-D1-K0$, $S2-D0-K0$, $S2-D1-K0$, $S1-D1-K1$, $S2-D0-K1$, or $S2-D1-K1$.

Input. The input contains the assigned values of S , D , K , and R , together with the core scene, subjects and objects, core state or interaction, scene context, reasoning clues, hidden expected outcome, core evaluation focus, and the original prompt skeleton.

General Enrichment Principles. Preserve the central scene, subjects, objects, and event defined in the original prompt skeleton. Do not replace the task with a different scenario merely to make it easier, harder, or more visually elaborate. Every added requirement must materially contribute to video-quality assessment and must be directly observable in the generated video. Avoid irrelevant decorative details and do not create artificial complexity through excessive adjectives or arbitrary visual constraints.

The enriched prompt must remain natural, coherent, and compatible with the short durations commonly supported by mainstream video generation models. All required content should be realizable within one continuous and logically coherent visual event. Avoid tasks that require numerous

scene transitions, extended narratives, or information external to the video. The public generation prompt must not contain metadata or terms such as “evaluation,” “benchmark,” “rubric,” “model capability,” $S1$, $D1$, or $K1$.

Scene-Complexity Enrichment. When $S = 1$, clearly specify the roles of the principal subjects and objects and introduce observable spatial constraints, such as their relative positions, orientations, distances, foreground-background relationships, or compositional arrangement. The scene should exhibit a clear but not excessively crowded organization. For dynamic tasks, the identities and roles of all interacting subjects or objects must remain visually distinguishable throughout the event.

When $S = 2$, establish a structured and visually interpretable organization involving multiple subjects or objects. This may include explicit foreground, middle-ground, and background layers, spatial groupings, relative positions, and visual hierarchies. For scenes involving occlusion, intersecting trajectories, or visually similar entities, add requirements concerning identity preservation, spatial continuity, and stable group organization. Scene complexity should arise from coherent multi-object composition rather than arbitrary accumulation of unrelated elements.

Dynamic-Interaction Enrichment. When $D = 0$, preserve the static or nearly static nature of the task and do not introduce unnecessary complex motion. Enrich the prompt primarily through subject poses, object shapes, spatial arrangement, material appearance, and static consistency. Minor movements may be included only when they improve visual naturalness without changing the fundamental static character of the task.

When $D = 1$, specify the observable initial state, key intermediate stages, and final state of the required action or interaction. Clarify the relevant motion directions, trajectories, temporal order, contact relationships, changes in velocity, and causal responses among subjects or objects. The complete process should remain temporally continuous and visually coherent. The generated video should avoid abrupt appearances or disappearances, identity exchanges, unintended penetrations, implausible deformations, or discontinuous state changes. For tasks involving collision, compression, penetration, cutting, material transformation, chemical reaction, or coordinated group motion, describe the most important intermediate stages needed to evaluate whether the process has been correctly realized.

Domain-Knowledge Enrichment. When $K = 0$, use objects, actions, and scene logic that can be understood through everyday knowledge. Emphasize natural motion, plausible interaction, clear composition, and visual coherence. Do not introduce unnecessary scientific terminology, experimental conditions, or specialized professional conventions.

When $K = 1$, identify the relevant professional or scientific domain and specify the key objects, operations, mechanisms, phenomena, or visual conventions that must be correctly depicted. Professional requirements must be translated into observable visual evidence rather than expressed through vague terms such as “professional,” “scientific,” or “accurate.”

For physical or material processes, specify the relevant material properties, applied forces, motion or deformation mechanisms, and expected observable responses. For chemical experiments, identify the relevant substances, operations, and visible outcomes, such as color changes, precipitation,

bubbling, crystallization, phase transitions, or other externally observable phenomena. For mechanics, air-pressure, or optical experiments, specify the apparatus, mode of operation, and expected visible result. For sports, dance, or professional performances, specify the action or technique, key poses, temporal sequence, coordination among participants, and any essential professional filming requirements. For typography, product presentation, aesthetic composition, or other professional static scenes, specify the relevant layout, text appearance, materials, poses, lighting, and compositional conventions.

Combination-Specific Requirements. For an $S1-D1-K0$ task, construct an everyday dynamic interaction involving multiple distinguishable subjects or objects. Emphasize action coordination, interaction order, contact relationships, motion continuity, and plausible causal responses without introducing unnecessary professional knowledge.

For an $S2-D0-K0$ task, construct a complex everyday static scene containing multiple subjects or objects. Emphasize spatial hierarchy, organized subject distribution, natural poses, object-level consistency, and overall compositional plausibility.

For an $S2-D1-K0$ task, construct a complex and highly dynamic everyday scene involving multiple subjects, simultaneous actions, or coordinated interactions. Emphasize group motion, concurrent object interactions, identity preservation, coherent trajectories, temporal continuity, and dynamically stable composition.

For an $S1-D1-K1$ task, construct a professional dynamic process involving several subjects or objects. Emphasize the professional operation, reaction, or interaction mechanism, its key intermediate stages, and the expected observable phenomenon.

For an $S2-D0-K1$ task, construct a complex professional static scene containing numerous subjects or objects. Emphasize the standardized arrangement of professional objects, subject poses, structural relationships, material depiction, visual hierarchy, and domain-appropriate composition.

For an $S2-D1-K1$ task, construct a complex, highly dynamic, and knowledge-intensive professional scene involving multiple subjects or objects. Emphasize professional actions, group coordination, multi-object interactions, physical or material consistency, and the complete temporal evolution of the intended event.

Reasoning-Task Processing. When $R = 0$, the final public prompt may explicitly describe the required action, process, phenomenon, and expected outcome. All principal task requirements may be directly stated.

When $R = 1$, the final public prompt may contain only the initial conditions, contextual information, and necessary reasoning clues. It must not explicitly state the hidden expected outcome, the name of the target phenomenon when that name directly reveals the answer, or the final event. Nevertheless, the clues must be sufficient for the expected outcome to be uniquely inferred through commonsense, physical principles, or relevant domain knowledge. The complete expected process and result must be retained in the internal task record and rubric information rather than disclosed in the generation prompt. If the available clues permit two or more substantially different but equally plausible outcomes, revise the clues until the intended outcome becomes uniquely inferable.

Output Format. Return the result strictly using the follow-

ing fields.

Task Level: Specify the assigned values of S , D , K , and R .

Final Video Generation Prompt: Produce one natural, complete, and directly usable text-to-video generation prompt. The prompt should be written in English and should integrate all necessary spatial, compositional, dynamic, interaction, and domain-specific requirements. For an $R1$ task, include only the permitted clues and do not disclose the hidden expected outcome.

Added Spatial and Compositional Constraints: State the newly introduced requirements concerning the principal subjects and objects, their identities, relative positions, orientations, spatial layers, grouping relationships, and overall composition.

Added Motion and Interaction Constraints: State the newly introduced requirements concerning action order, motion trajectories, contact relationships, temporal continuity, causal responses, and state transitions. For a $D0$ task, describe the relevant static-state and consistency requirements instead.

Added Domain-Specific Constraints: State the professional objects, operations, mechanisms, phenomena, material properties, or visual conventions that must be accurately presented. For a $K0$ task, output “No additional domain-specific constraints.”

Hidden Expected Process and Outcome: Describe the complete correct progression of the event and its intended final state. This information is intended for task filtering and rubric construction and must not be exposed in the public prompt for an $R1$ task.

Likely Failure Modes: Identify the principal errors that would reveal insufficient model capability, such as missing subjects, incorrect object counts, incoherent composition, unstable identities, discontinuous actions, implausible interactions, unintended deformation, incorrect physical behavior, inaccurate professional phenomena, or omitted task-specific details.

Candidate Rubric Criteria: Convert the task requirements into independently verifiable criteria. Distinguish the primary criteria, which directly determine whether the core task has been correctly completed, from the basic criteria, which assess non-primary object consistency, visual fidelity, aesthetic quality, and naturalness.

Self-Verification: Verify whether the final prompt conforms to the specified S , D , K , and R levels; whether every principal requirement can be assessed from visible evidence in the video; whether an $R1$ prompt conceals the expected result while still allowing it to be uniquely inferred; whether the complete event can reasonably occur within a short video; and whether any irrelevant details have been introduced. If any condition is not satisfied, revise the task before returning the final result.

Primary Criteria Generation Prompt

You are an expert evaluator for AI video generation models. Your task is to construct the primary evaluation criteria for a task-specific pairwise preference rubric. The primary criteria should focus on whether the generated video accurately satisfies the core requirements of the task and should capture the most important task-dependent factors that distinguish strong and weak generations.

General Requirements: - Generate 3-6 concise and independent primary criteria. - Each criterion must correspond to an observable property that can be verified from the generated video. - The criteria should prioritize task-specific requirements rather than generic video quality. - Do not include generic criteria such as visual fidelity, aesthetic appeal, naturalness, overall realism, or image quality, as these belong to secondary/basic criteria. - Avoid vague descriptions such as "high quality", "professional", or "good consistency". Instead, describe concrete visual evidence that should appear in the generated video. - The criteria should be suitable for pairwise comparison between two generated videos.

The task type can be either Text-to-Video (T2V) or Image-to-Video (I2V). Follow the corresponding rules below.

For Text-to-Video (T2V) Tasks:

Input information includes: - Original task prompt. - Enriched prompt details generated during task construction, including object quantities, spatial constraints, motion details, interaction processes, and domain-specific requirements. - Capability dimensions, including scene complexity (S), dynamic interaction complexity (D), and knowledge requirement (K). - Reasoning-task indicator.

Generate primary criteria by extracting and reorganizing the essential requirements from the original prompt and enriched details, especially the details.

Specifically: - For scene-complexity requirements, evaluate whether important subjects and objects, their quantities, spatial relationships, composition structures, and scene organization are correctly realized. - For dynamic-interaction requirements, evaluate whether the intended actions, motion patterns, interaction processes, temporal evolution, and causal relationships are correctly completed. - For knowledge-intensive tasks, evaluate whether domain-specific phenomena, mechanisms, materials, operations, or professional details are accurately represented. - For reasoning-based tasks, the prompt only provides clues rather than explicitly describing the expected outcome. Generate criteria based on the intended phenomenon or event recorded in the task construction information, and evaluate whether the model correctly infers and presents the expected content.

The criteria should mainly measure whether the generated video fulfills the core task requirements and realizes the most discriminative details among advanced AIVGMs.

For Image-to-Video (I2V) Tasks:

Input information includes: - Reference image(s). - Task prompt. - Task type (KFT or FLT). - Reasoning-task indicator. - For reasoning-based tasks, the expected phenomenon or event recorded in the task construction information.

For non-reasoning I2V tasks: - Directly reuse the task

prompt as the basis of the primary criteria, since the prompt explicitly describes the required event, interaction, or transformation. - Organize the explicit requirements in the prompt into concise and independently verifiable criteria. - Do not introduce additional requirements that are not specified by the prompt or reference images. - Additionally consider the constraints imposed by the reference images: - For FLT tasks, evaluate whether the generated video preserves the provided first and last frames and produces a coherent transition process between them. - For KFT tasks, evaluate whether the generated video maintains consistency with the provided keyframe while generating a logically consistent subsequent event.

For reasoning-based I2V tasks: - The prompt only provides clues and does not explicitly reveal the expected outcome. - Do not directly reuse the prompt as the primary criteria. - Generate criteria according to the expected phenomenon or event recorded during task construction. - Describe the observable visual evidence required to determine whether the model correctly infers and realizes the intended event from the provided clues and reference images. - Ensure that the criteria evaluate both the inferred outcome and its consistency with the reference content.

For all I2V tasks: - Prioritize task compliance, reference-image consistency, event or transformation realization, interaction correctness, and physical or logical consistency. - Do not include generic visual quality criteria, which belong to secondary/basic criteria.

Output Format:

Return only a JSON object:

"primary-criteria": ["Criterion 1", "Criterion 2", "Criterion 3"]

Each criterion should be a single concise sentence describing an independently observable requirement.

A.4 Human Experiment Details

For each evaluation task, we conduct a pairwise human preference study over the candidate generated videos. Given a task t and its candidate model set $\mathcal{M}_t$, each comparison pair is denoted as (i, j) , where $i, j \in \mathcal{M}_t$ and $i \neq j$. For each pair, annotators watch the two videos side by side and choose one of three outcomes: the left video is better, the right video is better, or the two videos are similar.

Cold-start stage Each task begins with a fixed cold-start stage before adaptive scheduling. We construct a coverage-balanced subset of comparison pairs $\mathcal{B}_t$ from the full candidate pair set $\mathcal{P}_t$. The cold-start pairs are selected with a round-robin strategy so that every model receives an initial number of comparisons. This prevents the adaptive scheduler from starting with completely unobserved ratings.

In our implementation, each task uses four cold-start rounds. All annotations collected during the cold-start stage are replayed into the TrueSkill rating model and are therefore included when estimating the initial posterior distributions for adaptive scheduling.

TrueSkill rating model For each model m , we maintain a latent quality distribution:

$$\mu_m \sim \mathcal{N}(\mu_m, \sigma_m^2),$$

where μ_m represents the current estimated quality of the model, and σ_m represents the uncertainty of this estimate. All models are initialized as

$$\mu_m = 5, \quad \sigma_m = \frac{5}{6}.$$

After each human comparison, the rating distributions of the two involved models are updated according to the TrueSkill posterior update rule. If one video is preferred, the winner’s mean rating is increased and the loser’s mean rating is decreased, while the uncertainties of both models are reduced. If the two videos are judged to be similar, the comparison is treated as a draw. The draw behavior is controlled by a draw probability parameter p_{draw} .

Adaptive pair scheduling After the cold-start stage, comparison pairs are selected adaptively. For each unannotated pair (i, j) , we compute the following acquisition score:

$$A(i, j) = \lambda_\sigma(\sigma_i + \sigma_j) + \exp\left(-\frac{|\mu_i - \mu_j|}{\beta}\right) + \lambda_c C(i, j) - \lambda_r R(i, j)$$

where μ_i, μ_j and σ_i, σ_j are the current TrueSkill mean and uncertainty values of the two models. The next pair is selected as

$$(i^*, j^*) = \arg \max_{(i, j) \in \mathcal{P}_t \setminus \mathcal{A}_t} A(i, j),$$

where $\mathcal{A}_t$ denotes the set of already annotated pairs for task t .

In our implementation, the scheduling weights are set as

$$\lambda_\sigma = 1.2, \quad \lambda_c = 0.6, \quad \lambda_r = 0.8,$$

with

$$\beta = \frac{25}{6}.$$

The first term,

$$\lambda_\sigma(\sigma_i + \sigma_j),$$

prioritizes model pairs whose ratings remain uncertain. The second term,

$$\exp\left(-\frac{|\mu_i - \mu_j|}{\beta}\right),$$

prioritizes pairs whose current estimated qualities are close, since comparisons between similarly rated models are more informative for refining the final ranking. The coverage bonus $C(i, j)$ encourages the scheduler to select under-compared models, while the repeat penalty $R(i, j)$ discourages repeatedly showing the same model in consecutive comparisons. All previously annotated pairs are excluded from the candidate set, preventing duplicate comparisons.

Convergence criterion The adaptive process is checked after every 10 newly annotated adaptive pairs. At each checkpoint, we compare the current model ranking with the ranking from the previous checkpoint using Spearman’s rank correlation coefficient:

$$\rho = \text{SRCC}(\text{rank}_k, \text{rank}_{k-1}).$$

A task is considered converged when all three pairwise SRCC values among the most recent three rankings exceed 0.95, and the mean uncertainty satisfies:

$$\frac{1}{|\mathcal{M}_t|} \sum_{m \in \mathcal{M}_t} \sigma_m \leq 0.3.$$

Once a task satisfies the convergence condition, no additional pairs are scheduled for that task.

In short, the adaptive scheduler favors pairs that are still uncertain and whose current TrueSkill means are close. This makes the annotation process focus on comparisons that are most useful for stabilizing the final model ranking.

A.5 Additional Experiments and Details for Leaderboards Construction

Due to space constraints in the main paper, we provide additional implementation details for the RAVEN-Eval Leaderboards below.

AIVGM Leaderboards We estimate the 95% confidence interval of each AIVGM score through stratified bootstrap resampling. Specifically, we perform 200 independent bootstrap rounds, each containing 2,000 model pairs sampled with replacement. The sampling budget is distributed approximately uniformly across all tasks to avoid overrepresenting particular tasks. To keep the key settings, for every sampled pair, we retain the complete set of judgments from all LMM judges rather than resampling individual judge outputs. We then rerun the same tie-aware optimization for each replicate to obtain a new score for every AIVGM. The 2.5th and 97.5th percentiles of the resulting score distribution are reported as the lower and upper confidence bounds, respectively. All other leaderboard construction details follow the main paper.

As for external online references, we provide the exact shared AIVGMs with AA and *Arena*. in Tab.6.

Judge Leaderboards To improve the robustness of the reported SRCC and KRCC values in the RAVEN-Eval-Judge Leaderboards, we **combine the estimate from the fully connected comparison graph (complete graph) with bootstrap-based estimates**. For each LMM judge, we first optimize one ranking from the complete graph and construct another 200 rankings using the bootstrap protocol described above. We compute the correlation value (SRCC and KRCC) between each LMM-based ranking and the reference ranking (including human or external rankings), and first average the corresponding correlations over the 200 bootstrap replicates. The final reported value (the **bootstrap-stabilized SRCC / KRCC**) is then obtained by averaging this bootstrap mean with the correlation value derived from the complete graph, assigning each component a weight of 1/2. **The correlation results in Tabs. 2 and 4 of the main paper follow**

this bootstrap-stabilized protocol, which reduces sensitivity to a single complete-graph optimization.

For LMMs that provide an explicit reasoning-effort option, we enable reasoning at the lowest supported level; otherwise, we use the default inference configuration. The 8 sampled frames from each video are provided at their native resolution.

Additional Leaderboards We separately report another 6 AIVGM leaderboards for *K1* T2V tasks, *S2* T2V tasks, *D1* T2V task, all reasoning-based T2V and I2V tasks, and the aggregated I2V tasks, together with an overall leaderboard aggregating the T2V, KFT, and FLT categories. If a task is incompatible with a particular model, it is excluded from the computation of that model’s final mean and variance. The resulting rankings are presented in Tab. 7.

Optimization Settings and Ablations The optimization hyperparameters are fixed across task categories and summarized in Tab.5. We further evaluate several variants of the proposed score-estimation method.

First, we remove explicit tie modeling. Both confidently identified ties and uncertain judgments are represented by assigning equal probability mass to the two directional outcomes. We retain the confidence-aware (dynamic λ_p) setting. Under this setting, the optimization reduces to a standard Bradley-Terry maximum likelihood formulation.

Second, we remove confidence-aware weighting and consider two alternatives. In the *full-trust* ($\lambda_p = 1$) setting, confident and uncertain judgments contribute equally to optimization. In the *confident-only* setting, uncertain judgments are discarded, while confident judgments retain full weight. The results are reported in Tab.9.

The standard Bradley-Terry formulation performs relatively worse than our modified Davidson tie-aware estimator, confirming the importance of explicitly modeling ties. Assigning full trust to uncertain judgments also causes a clear performance drop. By contrast, the confident-only variant performs close to, but slightly below, our confidence-aware setting. These observations suggest that uncertain outputs mainly reflect unstable or unresolved directional preferences rather than genuine judgments that two videos have similar quality. Treating them as fully reliable therefore introduces noise, whereas adaptive downweighting preserves limited useful evidence without allowing uncertain outcomes to dominate score estimation.

Effects of the Number of LMM Judges We observe an apparent difference between the main-paper results: individual LMM judges generally achieve stronger agreement with the human reference on FLTs than on KFTs in Tab. 4 in main paper, whereas the three-judge evaluation in Tab. 2 performs better on KFTs than on FLTs. This discrepancy may partly result from the random composition of the 10 annotated KFTs and 10 annotated FLTs. More importantly, we assume that the performance gain from combining multiple judges may appear substantially larger than a single judge.

To further examine the effects of the Number of LMM judges, we conduct an additional ablation using all 50 human-annotated tasks and construct a unified ranking without sepa-

rating T2V and I2V tasks. For AIVGMs that do not support FLTs, unavailable FLT scores are omitted during aggregation. As in the main evaluation, the final score is computed from the mean task-level score with a standard-deviation penalty. We estimate rankings using $M = 1$, $M = 2$, and $M = 3$ judges, evaluate every possible judge combination under each ensemble size, and report bootstrap-stabilized SRCC values against the human reference. The results are presented in Tab 10.

We further select the top-7 LMM judges from the T2V Judge Leaderboard and evaluate ensemble sizes from $M = 1$ to $M = 7$. For each M , we enumerate all possible judge combination of size M and report their mean bootstrap-stabilized SRCC on all the 50 tasks with human reference. The resulting curve (Fig.4) shows a clear gain when increasing the ensemble size to three judges and a sustained upward trend as more capable judges are incorporated. These results indicate that, when the candidate judges are individually reliable, increasing the ensemble size generally improves overall evaluation accuracy and reduces sensitivity to any particular judge combination. **This finding also provides a plausible explanation for the stronger KFT performance observed under the three-judge setting in the main paper.**

More details about the anchor-based model insertion paradigm The **experimental complexity reduction** shown in the main paper ($K = 1$:34.2%, $K = 2$:21.1%) is derived from below. The theoretical reduction is based solely on the number of pairwise comparisons is 31.6% for $K = 1$ and 18.4% for $K = 2$. Specifically, a fully connected graph over 20 models contains $\binom{20}{2} = 190$ model pairs. After retaining the fully connected graph among the 15 existing models, the anchor-based strategy requires $105 + 5 \times 5 = 130$ comparisons for $K = 1$ and $105 + 5 \times 10 = 155$ comparisons for $K = 2$. The corresponding reductions are therefore $(190 - 130)/190 = 31.6\%$ and $(190 - 155)/190 = 18.4\%$, respectively. In our implementation, we further account for the reduction in actual LMM invocation rounds. The anchor-based insertion procedure avoids additional LMM API calling rounds, which can be roughly treated as 5 pairwise-comparison-equivalent calls. After this implementation-level adjustment, the effective numbers of saved calls become $(190 - 130) + 5 = 65$ and $(190 - 155) + 5 = 40$, yielding effective cost reductions of $65/190 = 34.2\%$ for $K = 1$ and $40/190 = 21.1\%$ for $K = 2$. Thus, 31.6% and 18.4% are the reductions derived purely from graph sparsification, whereas 34.2% and 21.1% additionally reflect the saved LMM API calling rounds in the actual evaluation pipeline.

As additional models are incorporated, the relative reduction in pairwise evaluation cost is expected to become more pronounced, because the cost of fully connected evaluation grows quadratically with the model-pool size, whereas anchor-based insertion requires comparisons only against a limited anchor set. However, continued expansion of the model pool may weaken the reliability and representativeness of the existing anchors, particularly when newly added models alter the capability distribution or introduce previously uncovered generation characteristics. The anchor set should therefore be periodically refreshed by conducting full

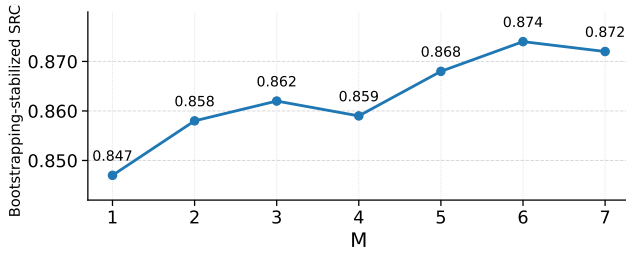


Figure 4: The trend of bootstrapping-stabilized SRCC as a function of M .

pairwise evaluation on selected updates or by dynamically expanding the comparison graph to identify new reliable anchors. Developing an adaptive anchor-maintenance strategy that balances evaluation efficiency against ranking reliability constitutes an important direction for future work.

Computational Analysis Computational Analysis of the whole estimation process is shown in Tab.8.

A.6 Benchmark Cases

T2V Prompts and Task-Specific Rubrics Figures 5–10 present the complete prompts and task-specific rubrics of the 48 selected T2V benchmark cases. The cases are organized according to the six retained configurations of scene complexity S , dynamic interaction complexity D , and knowledge requirement K . Ten cases are additionally marked as reasoning tasks. To avoid redundant text without omitting any evaluation criterion, the shared basic rubric is presented once at the top of each page and applies to all cases on that page. Each individual case retains its complete prompt, primary rubric, and additional task-specific criteria.

T2V Case Examples We present representative T2V case examples for the six benchmark difficulty configurations and an additional reasoning-focused case in Figs. 11–17. Each case includes the complete English prompt, the task-specific primary rubric, and the outputs of six evaluated video generation models. For every model, eight frames are sampled at fixed relative temporal positions and arranged in a 2×4 grid. The model selection varies across cases and collectively covers all 20 evaluated video generation models.

I2V case Examples We provide representative case examples for the two I2V task categories, i.e., first–last-frame transition (FLT) and keyframe continuation (KFT), as shown in Figs. 18–25. For each category, we include reasoning and non-reasoning cases collected from both YouTube and ASMR videos. Each case presents six different evaluated models, with eight uniformly sampled output frames arranged in a 2×4 grid for each model. For FLT, only the first and last input frames are displayed, whereas KFT shows only the starting keyframe. The model selection varies across cases and collectively covers all 20 evaluated video generation models.

Prompt Word Clouds Figure 30 summarizes the lexical distributions of the prompts used in the three evaluation tasks.

We aggregate all available prompts for T2V, FLT, and KFT, translate the original Chinese text into English using an off-line machine-translation model, and then apply lowercasing, tokenization, stop-word removal, and basic singular-form normalization. The resulting corpora contain 150 T2V prompts, 50 FLT prompts, and 50 KFT prompts. Word size is proportional to corpus-level frequency after preprocessing. The T2V prompts emphasize motion, interaction, spatial relationships, stability, and material or physical properties, whereas the FLT and KFT prompts contain more concrete objects, materials, and state changes associated with temporally constrained video generation.

Failure-Case Analysis We conduct a failure-case analysis for each task category. Since the RAVEN-Eval AIVGM Leaderboards exhibit strong correlations with established external human-annotation-based leaderboards, we use the corresponding category-level RAVEN-Eval ranking as a proxy reference. This choice is necessary because the external platforms do not provide separate rankings for KFT and FLT. T2V and KFT tasks evaluated by at least 18 AIVGMs and FLT tasks evaluated by at least 14 AIVGMs. For each eligible task, we compute the SRCC between its task-specific model ranking and the corresponding category-level ranking over their shared models. We then select the three tasks with the lowest SRCC in each category, yielding nine highly probable failure cases, as shown in Figs. 26,27,28. Although these tasks may contain task-specific biases that produce unusually divergent rankings, they remain informative for diagnosing systematic weaknesses in the evaluation framework. We further employ LMM-assisted qualitative analysis to identify their underlying failure modes.

Our preliminary analysis reveals distinct patterns across task categories. For T2V, the primary failure mode arises from an **imbalance between fine-grained requirement satisfaction and holistic scene quality**. In several inconsistent cases, the evaluation places excessive emphasis on localized details, such as object counts or specific poses, while underweighting the superior overall composition, coherence, and visual appeal produced by some strong models. This imbalance can lead to disagreement between the automated ranking and the proxy human-oriented reference. For KFT and FLT, **the main difficulties involve highly specialized and complex scenarios, as well as the assessment of temporally continuous actions and fine-grained motion evolution**. Because the current judging pipeline relies primarily on sparsely sampled video frames, it has an inherent limitation in evaluating motion continuity, action completeness, and intermediate state transitions, which can reduce ranking accuracy. These findings motivate two directions for future work: better balancing holistic scene quality against local task-specific details in rubric design, and augmenting the primary LMM judge with complementary temporal-analysis models or external tools for tasks involving complex dynamics and specialized phenomena.

Win-rate Heatmaps We also provide the pairwise win-rate heatmaps for the T2V, KFT, and FLT tasks, computed from all 3 LMM judge results. It is shown in Fig.29.

Hyperparameter	Value	Definition
Optimizer	L-BFGS-B	Quasi-Newton optimizer used to minimize the Davidson negative log-likelihood.
Learning rate	N/A	No fixed learning rate is used; step sizes are selected automatically by L-BFGS-B line search.
maxiter	1,000	Maximum number of optimizer iterations allowed for each task-level fit.
ftol	10^{-11}	Convergence tolerance; optimization may stop when the relative improvement in objective value becomes sufficiently small.
gtol	10^{-6}	Gradient convergence tolerance; optimization may stop when the projected gradient norm is sufficiently small.
Initialization of s_i	0	Initial skill score for every model in each task-level Davidson fit.
Initialization of $\log \eta$	0	Initial log tie parameter; equivalent to initializing $\eta = \exp(\log \eta) = 1$.
α	10^{-3}	L2 regularization coefficient on model skills, applied as $\alpha \sum_i s_i^2$.
λ_g	10.0	An implementation-level gauge-fixing penalty corresponding to the term $\sum_{i=1}^N s_i = 0$ in the main paper.
β	10^{-4}	L2 regularization coefficient on the log tie parameter, applied as $\beta(\log \eta)^2$.
η	learned	Davidson tie parameter controlling tie probability, parameterized as $\eta = \exp(\log \eta)$ to ensure positivity.
$\bar{s}$	task mean	Mean task-level skill, $\bar{s} = \frac{1}{N} \sum_i s_i$, used in the gauge regularization term.
Bootstrap samples per round	2000	Number of model pairs sampled with replacement in each bootstrap round.
Bootstrap rounds	200	Number of independent bootstrap leaderboard reconstructions used for confidence-interval estimation.
CI level	95%	Percentile interval computed from the 2.5th and 97.5th percentiles of bootstrap Elo samples.

Table 5: Hyperparameters used in the Davidson tie-aware Bradley–Terry optimization and leaderboard aggregation.

Task set	Reference	Shared models	Count
T2V	AA	seedance2, grok_image_video, kling3-pro, happyhorsev1, kling3-omni, veo3.1-fast, veo3.1-lite, pixversev6, kling3-std, wan2.6, pixversev5.5, seedance1.5-pro, ltx2.3-fast, wan2.7, ltx2.3-pro, hailuo2.3, ltx2-pro, ltx2-fast	18
T2V	Arena.AI	seedance2, grok_image_video, happyhorsev1, wan2.6, pixversev5.5, seedance1.5-pro, wan2.7, hailuo2.3, pixversev6, veo3.1-lite, veo3.1-fast, kling3.0-pro, kling-3.0-omni, kling-3.0-std	14
KFT	AA	grok-image-video, seedance2, kling3.0-pro, wan2.7, kling-3.0-omni, kling-3.0-std, happyhorsev1, pixversev6, wan2.6, veo3.1-lite, seedance1.5-pro, hailuo2.3, veo3.1-fast, pixversev5.5, ltx2.3-pro, ltx2-fast, ltx2.3-fast	18
KFT	Arena.AI	seedance2, grok_image_video, happyhorsev1, wan2.6, pixversev5.5, seedance1.5-pro, wan2.7, hailuo2.3, pixversev6, veo3.1-lite, veo3.1-fast, kling3.0-pro, kling-3.0-omni, kling-3.0-std	14
FLT	AA	seedance2, kling3.0-pro, wan2.7, kling-3.0-omni, kling-3.0-std, pixversev6, veo3.1-lite, seedance1.5-pro, veo3.1-fast, pixversev5.5, ltx2.3-pro, ltx2.3-fast	12
FLT	Arena.AI	seedance2, pixversev6, kling3.0-pro, wan2.7, seedance1.5-pro, veo3.1-lite, veo3.1-fast, kling-3.0-omni, pixversev5.5, kling-3.0-std	10

Table 6: Overlapping models between our leaderboard and external reference leaderboards for alignment computation.

Category	Reasoning		S2		D1		K1		I2V overall		Overall	
AIVGM	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score
Seedance 2	1	999.7	2	854.6	1	852.6	1	928.8	1	978.9	1	904.1
Seedance 2 Fast	2	930.7	6	723.5	6	691.6	10	652.4	5	809.1	4	774.5
Grok-Imagine-Video	3	834.5	5	759.1	2	799.0	4	803.3	2	947.0	2	857.1
Kling 3.0 Omni	4	752.7	8	699.0	5	693.2	2	865.8	7	787.0	5	762.3
PixVerse C1	5	741.3	9	557.1	8	600.0	12	579.5	3	840.6	7	718.0
Kling 3.0 Pro	6	718.4	12	485.0	4	750.0	8	697.4	4	825.6	3	799.6
Kling 3.0 Std	7	654.4	20	72.4	11	489.4	7	721.7	8	782.4	8	677.2
Wan 2.7	8	589.9	4	812.6	15	284.9	9	655.0	6	791.9	10	573.8
Veo 3.1 Fast	9	565.2	3	839.6	7	673.7	14	558.1	15	323.1	13	447.6
PixVerse V6	10	533.0	10	555.1	10	531.4	11	644.9	10	646.0	9	601.2
Veo 3.1 Lite	11	483.7	1	966.4	9	590.5	6	764.4	12	478.6	12	524.7
HappyHorse V1.0	12	414.3	7	718.8	3	768.8	5	797.2	9	690.1	6	730.7
Wan 2.6	13	413.6	14	377.7	13	468.5	13	565.3	11	638.3	11	556.1
PixVerse V5.5	14	270.6	15	297.9	12	469.6	18	285.0	16	294.2	15	357.3
Seedance 1.5 Pro	15	257.3	19	218.3	14	361.9	16	456.4	13	421.7	14	401.8
LTX 2.3 Pro	16	87.2	16	284.8	18	-31.4	15	468.9	17	-52.4	17	-61.4
LTX 2.0 Pro	17	38.6	13	437.7	17	-29.3	17	338.4	18	-323.2	18	-160.1
LTX 2.0 Fast	18	22.0	17	259.9	20	-117.8	19	15.4	19	-336.8	20	-232.4
LTX 2.3 Fast	19	-10.3	11	534.5	16	276.4	3	834.3	20	-382.9	19	-208.4
Hailuo 2.3	20	-259.9	18	220.0	19	-62.8	20	-169.9	14	400.1	16	128.7

Table 7: Sub-leaderboards display. Each leaderboard reports only rank and Elo-style score.

Item	Setting / Value	Description
CPU	2 × Intel Xeon Platinum 8558	96 physical cores / 192 threads in total.
Memory	2.0 TiB RAM	System memory available on the evaluation machine.
Operating mode	CPU-only	Pairwise likelihood optimization, Elo aggregation, and bootstrap sampling were run on CPU.
Python	3.11.10	Runtime environment used for the benchmark.
NumPy	2.2.6	Used for array operations, random bootstrap sampling, mean/std computation, and percentile CI computation.
SciPy	1.16.0	Used for task-level Davidson optimization via <code>scipy.optimize.minimize</code> .
Optimizer	L-BFGS-B	Quasi-Newton optimizer used for the Davidson tie-aware Bradley–Terry objective.
Task set	150 T2V + 50 KFT + 50 FLT	Combined benchmark set, 250 task-level pairwise optimization problems in total.
Bootstrap samples	2000	Number of resampled leaderboard estimates used for confidence interval estimation per round.
Random seed	20260631	Seed used for bootstrap resampling.
Task-level fits	250 / 250 successful	All task-level Davidson optimizations converged successfully in the measured run.
Cached-score leaderboard latency	8.15 s	Time for loading fitted task scores, computing the Elo leaderboard, and running 2000 bootstrap samples.
End-to-end optimization latency	60.86 s	Time for parsing pairwise JSONs, fitting all task-wise Davidson models, computing the Elo leaderboard, and running all bootstrap rounds samples.

Table 8: Computational setup and measured latency for producing the combined T2V/KFT/FLT leaderboard. The cached-score latency starts from previously fitted task-level Davidson scores, while the end-to-end latency includes re-fitting all task-level Davidson models from pairwise JSON files.

Category Setting	T2V			I2V			
	<i>Human</i>	<i>AA</i>	<i>Arena</i>	<i>H. (KFT)</i>	<i>H. (FLT)</i>	<i>AA</i>	<i>Arena</i>
<i>Standard BT-MLE</i>	0.855	0.815	0.801	0.871	0.813	0.843	0.703
<i>Confident-only</i>	0.867	0.825	0.812	0.894	0.817	0.846	0.712
<i>Full-trust</i>	0.832	0.799	0.787	0.884	0.785	0.830	0.690
<i>Ours</i>	0.872	0.835	0.810	0.903	0.821	0.851	0.714

Table 9: Ablation results (SRCC) under different optimization settings. The term “H.” is short for “Human.” The highest and second-highest values in each column are shown in bold and italics, respectively.

LMM Judges	Human Reference (50 Tasks)
<i>GPT-5.4</i>	0.864
<i>Gemini 3.1 FL</i>	0.817
<i>Claude Sonnet 4.6</i>	0.856
<i>GPT-5.4 + Gemini 3.1 FL</i>	0.837
<i>GPT-5.4 + Claude Sonnet 4.6</i>	0.870
<i>Claude Sonnet 4.6 + Gemini 3.1 FL</i>	0.848
<i>All</i>	0.874

Table 10: Ablation results (SRCC) on the effect of the number and composition of LMM judges. The highest and second-highest values are shown in bold and italics, respectively.

A.7 Limitations

Although this work provides detailed algorithmic analyses and extensive validation of its key components, several limitations remain. Due to the substantial costs of human annotation, proprietary high-performance LMM judges, and closed-source AIVGMs, we are unable to obtain human annotations for all tasks or evaluate the full set of 250 tasks with a broader pool of capable LMM judges. The current AIVGM evaluation is also limited to 20 models. Consequently, the evidence for the large-scale scalability of RAVEN-Eval and its potential for LMM judge model harnessing still need further research. Extending the benchmark to more generation models, judge ensembles, and human-validated tasks is therefore a central direction of our future work.

SHARED BASIC RUBRIC — applies to every task on this page (1) All static objects and background objects must remain visually consistent throughout the video. (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible. (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes. (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.	
01. T2V_L1_D1_K0_0001 Prompt. One child throws a yellow ball to another child, who catches it and steps slightly backward. Use a fixed camera. Emphasize direct interaction between a small number of subjects, preserve natural continuity across the throw and catch, and make the action order and rhythm of both participants clear. Primary rubric. (1) The ball travels toward the second child along a continuous trajectory. (2) The second child successfully catches the ball and reacts by stepping slightly backward. (3) The full sequence of throwing, flight, and catching is complete and continuous. (4) The spatial relationship between the two children and the outcome of the interaction remain clearly visible. Additional criteria. (1) The visual treatment should emphasize direct interaction between a small number of subjects. (2) The yellow ball should serve as a visually salient interaction focus in the two-person throwing-and-catching sequence.	06. T2V_L1_D1_K0_0010 Prompt. Two people jointly lift a small cardboard box and place it on a nearby table. Their hands, the box, and their body movements should remain continuous and natural. Use a fixed camera, emphasize dynamic interaction between a small number of subjects, and keep the main motion path stable without sudden displacement of the participants or box. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows two people jointly moving a small cardboard box. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should emphasize dynamic interaction between a small number of subjects.
02. T2V_L1_D1_K0_0003 Prompt. Two cyclists approach each other from opposite directions on a path and pass one another. At the moment they meet, both make a slight handlebar adjustment. Keep the background simple and the camera fixed. Emphasize the path relationship and dynamic connection between the two subjects, with a clear action order and natural rhythm. Primary rubric. (1) The two cyclists approach from opposite directions. (2) Both cyclists maintain continuous motion before they meet. (3) Both make a slight handlebar adjustment at the moment of passing. (4) They successfully pass each other, and their path relationship remains clear. (5) Their motion trajectories remain continuous and stable throughout. Additional criteria. (1) The visual treatment should emphasize the path relationship and dynamic connection between the two subjects. (2) A simple background should foreground the subjects and preserve a restrained visual quality.	07. T2V_L1_D1_K0_0012 Prompt. One person steadies a short ladder and helps another person slowly step down from the final rung while the second person adjusts their footing. Use a fixed camera, emphasize the interaction created by a shared action rhythm, and keep the main motion path stable without sudden displacement. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows one person steadying the ladder while another descends. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should emphasize the interaction created by the subjects' shared action rhythm.
03. T2V_L1_D1_K0_0004 Prompt. One person opens an umbrella and then walks forward with another person. The second person moves under the umbrella and adjusts their pace. Use a fixed camera. Emphasize the interaction created when a small number of subjects share an action rhythm, and keep the action order and rhythmic transitions clear and natural. Primary rubric. (1) The first person completes the umbrella-opening action before walking. (2) The two people walk forward together. (3) The second person moves under the umbrella and adjusts their pace. (4) The shared-umbrella relationship remains clearly visible. Additional criteria. (1) The two participants' motion rhythms and positional changes should be natural and continuous. (2) The visual treatment should emphasize the interaction created by the subjects' shared action rhythm. (3) The background atmosphere, supporting details, and visual rhythm of the moving scene should remain consistent with the prompt.	08. T2V_L1_D1_K0_0013 Prompt. One person tilts a teapot to pour tea into another person's cup; the second person steadies the cup and then steps slightly backward. Use a fixed camera, emphasize direct interaction between the two subjects, preserve natural continuity before and after pouring, and fully show the key contact point and transfer moment. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows one person pouring tea for another. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should emphasize direct interaction between the two subjects.
04. T2V_L1_D1_K0_0007 Prompt. Two joggers approach in parallel on a path and exchange a relay baton. At the handoff, both slightly adjust the direction of their arms. Keep the background simple and the camera fixed. Emphasize their path relationship and dynamic connection, maintain reasonable changes in distance, and make the interaction outcome clearly visible. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on two-subject coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows two joggers exchanging a relay baton. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should show the subjects' path relationship and dynamic connection, preserve reasonable distance changes, and clearly present the outcome.	09. T2V_L1_D1_K0_0019 Prompt. Two people stand on opposite sides of a table and gently pull a tablecloth flat at the same time. During the pull, both slightly adjust the direction of their arms. Keep the background simple and the camera fixed. Emphasize their path relationship and dynamic connection, including natural reactions and subtle posture adjustments. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows two people jointly pulling a tablecloth flat. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should emphasize the subjects' path relationship and dynamic connection, including natural reactions and subtle posture adjustments.
05. T2V_L1_D1_K0_0009 Prompt. A person leads a small dog around a yellow traffic cone. After the dog turns, the owner steps slightly backward. Use a fixed camera, emphasize direct interaction between the two subjects, preserve natural continuity in the guiding action, and keep the main motion path stable without sudden displacement of either subject. Actions should be continuous, directionally clear, and moderate in amplitude, without abrupt jumps. Use a stable medium shot with clear action paths and a readable interaction, focusing on coordination, credible contact, and motion continuity. Primary rubric. (1) The main scene shows an owner guiding a small dog around an obstacle. (2) Two principal interacting subjects are present. (3) The action path is continuous, with clear start and end states. Additional criteria. (1) The distance or contact relationship between the subjects is reasonable. (2) The camera remains stable and the interacting subjects do not drift. (3) The visual treatment should emphasize direct interaction between the two subjects.	

Figure 5: Complete T2V benchmark cases for $S=1$, $D=1$, and $K=0$: moderate semantic complexity, rich detail, and no external knowledge requirement.

SHARED BASIC RUBRIC — applies to every task on this page (1) All static objects and background objects must remain visually consistent throughout the video. (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible. (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes. (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.	
10. T2V_L1_D1_K1_0001 Prompt. Two investigators in long trench coats pass one another in a neon-lit alley on a rainy night. One raises a hand against the rain while the other looks back. Use a Blade Runner-style cyberpunk film-noir aesthetic with wet-ground reflections, cyan-purple neon, high-contrast shadows, layered mist, and a low camera angle. Motion should remain continuous and natural, with emphasis on cinematographic style and character-path relationships. Primary rubric. (1) Two investigators wearing long trench coats are visible. (2) They pass each other in a neon-lit alley on a rainy night with continuous motion. (3) The overall scene presents a Blade Runner-style cyberpunk film-noir aesthetic. Additional criteria. (1) Wet-ground reflections, mist, and high-contrast shadows are clearly visible. (2) The low camera angle and cyan-purple neon palette remain consistent. (3) The stylistic treatment includes wet-ground reflections, cyan-purple neon, high-contrast shadows, layered mist, and a low camera angle. (4) The visual treatment should convey cinematic photography and the characters' path relationship.	14. T2V_L1_D1_K1_0006 [REASONING] Prompt. A hydraulic press slowly descends toward a rubber duck, with the press head, duck, and workbench visible together. The complete process should show the press head moving continuously downward, making contact with the target, and producing subsequent changes that follow the natural response of a material under force. The interaction path, final state, and professional scene logic should remain clear and explainable. Primary rubric. (1) After the press head contacts the rubber duck, the contact area should flatten first and deform visibly along the vertical force direction. (2) The duck should exhibit elastic-material behavior, such as body compression, lateral bulging, or slight displacement of the head or beak. (3) As the press continues downward, the deformation should deepen gradually; the duck should not suddenly shatter like a brittle material. (4) Contact among the press head, rubber duck, and workbench should remain stable, without interpenetration, floating, or an incorrect force direction. (5) The final state should show a physically reasonable elastic deformation caused by sustained compression. Additional criteria. None specified.
11. T2V_L1_D1_K1_0002 Prompt. Two modified off-road vehicles race side by side along a dusty section near a desert canyon. The lead vehicle makes a slight drift while the following vehicle avoids loose rocks. Use a Mad Max: Fury Road-style desert action-film aesthetic with high-contrast orange-yellow color, a handheld chase-camera feeling, airborne dust, exaggerated mechanical silhouettes, and a strong sense of speed. Emphasize the dynamic relationship between the vehicles and the cinematographic rhythm. Primary rubric. (1) Two modified off-road vehicles are visible. (2) The vehicles move side by side through a desert canyon or dusty road section. (3) The scene presents a Mad Max: Fury Road-style desert action-film aesthetic. Additional criteria. (1) The high-contrast orange-yellow palette and dust particles are prominent. (2) The changing front-to-back positions and sense of speed are clear. (3) The style includes a handheld chase-camera feeling, dust particles, exaggerated mechanical silhouettes, and strong speed. (4) The visual treatment should convey the dynamic relationship between the two vehicles and the cinematographic rhythm.	15. T2V_L1_D1_K1_0009 [REASONING] Prompt. A beam of light travels from a light source through a prism and onto a screen behind it. The frame contains at least the light source, prism, and screen. The light should interact continuously with these objects, and the resulting effect should follow common optical laws. Emphasize the professional dynamic relationship among a small number of objects while keeping the interaction path, final state, and scene logic clear and explainable. Primary rubric. (1) Before entering the prism, the light beam should follow a relatively concentrated propagation path. (2) After passing through the prism, the screen should show a positional change, displaced light spot, or dispersion effect caused by refraction. (3) If dispersion appears, the light on the screen should separate into a continuous or nearly continuous multicolored band rather than unrelated colors flashing randomly. (4) The light source, prism, optical path, and screen response should remain spatially aligned and causally connected. (5) The optical effect should remain stable with the light path and must not jump abruptly or appear unrelated to the prism's position. Additional criteria. None specified.
12. T2V_L1_D1_K1_0003 Prompt. One astronaut slowly walks toward a giant black monolith while a second astronaut stands beside a rover and observes. Use a 2001: A Space Odyssey-style austere science-fiction aesthetic with symmetrical composition, minimalist space, slow movement, cool white light, and extensive silent negative space. Motion should be subtle but continuous, emphasizing cinematic composition and a solemn atmosphere. Primary rubric. (1) The frame contains two astronauts and one giant black monolith. (2) One astronaut slowly approaches the monolith while the other observes beside the rover. (3) The scene presents an austere, minimalist science-fiction aesthetic in the style of 2001: A Space Odyssey. Additional criteria. (1) Symmetrical composition, cool white light, and extensive negative space are evident. (2) The movement remains slow and the spatial relationships remain stable. (3) The style includes symmetrical composition, minimalist space, slow movement, cool white light, and extensive silent negative space. (4) The visual treatment should convey cinematic composition and a solemn atmosphere.	16. T2V_L1_D1_K1_0008 [REASONING] Prompt. A double-pendulum-ball experimental apparatus shows both balls and the support structure in the same frame. During the experiment, one ball is released and contacts the other. The interaction and subsequent changes should follow common mechanical laws. Emphasize professional interaction between a small number of objects while keeping the interaction path, final state, and professional scene logic clear and explainable. Primary rubric. (1) The released pendulum ball should swing along an arc toward the other ball, following gravity-driven pendulum motion. (2) A clear collision moment or contact position should be visible when the two balls meet. (3) After the collision, the struck ball should begin moving, while the released ball slows, pauses, or swings backward. (4) The changes in the two balls' motion should demonstrate momentum or energy transfer rather than unrelated simultaneous motion. (5) The balls, strings, and support should remain physically constrained: the balls must not detach, pass through each other, or suddenly change size. Additional criteria. None specified.
13. T2V_L1_D1_K1_0005 Prompt. An explorer and an assistant slowly step backward in misty forest ruins as a huge shadow sweeps across the foreground; both look upward at the same time. Use a Jurassic Park-style Spielberg adventure-film aesthetic with a low-angle upward shot, warm flashlight illumination, layered mist, and staging that first builds suspense and then reveals a silhouette. Emphasize both characters' reactions and the adventure-film atmosphere. Primary rubric. (1) An explorer and an assistant are visible. (2) They step backward and look upward in misty forest ruins. (3) The scene presents a Jurassic Park-style adventure-film aesthetic. Additional criteria. (1) The low-angle shot, warm flashlight illumination, and layered mist are clearly visible. (2) A large shadow or silhouette creates suspense without making the scene confusing. (3) The style includes low-angle framing, warm flashlight light, layered mist, and suspense-first silhouette revelation. (4) The visual treatment should convey both characters' reactions and an adventure-film atmosphere.	17. T2V_L1_D1_K1_0010 [REASONING] Prompt. A pneumatic experimental apparatus shows a main container, a pressure gauge, and a flexible component in the same frame. During the experiment, the internal state of the apparatus changes gradually and the components remain causally connected. Subsequent changes should follow common pressure behavior. Emphasize the interaction logic among a small number of professional objects while keeping the interaction path, final state, and professional scene logic clear and explainable. Primary rubric. (1) During pressurization or depressurization, the pressure-gauge needle, reading, or indicator state should change visibly with the internal pressure. (2) The flexible component should respond correspondingly by expanding, contracting, bulging, collapsing, or recovering its shape. (3) The pressure-gauge change and flexible-component response should be temporally correlated rather than contradictory or causally unrelated. (4) The main container, pressure gauge, and flexible component should maintain a clear physical connection or functional relationship. (5) The full process should form a continuous physical chain from pressure change to a visible mechanical or deformation response. Additional criteria. None specified.

Figure 6: Complete T2V benchmark cases for $S=1$, $D=1$, and $K=1$: moderate semantic complexity, rich detail, and an external knowledge requirement. Reasoning cases are explicitly marked in the figure.

SHARED BASIC RUBRIC — applies to every task on this page (1) All static objects and background objects must remain visually consistent throughout the video. (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible. (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes. (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.	
18. T2V_L2_D0_K0_0001 Prompt. A crowded bookstore interior contains at least 6 people, 4 shelving areas, and more than 30 visible books. People remain reading or standing, with only very small page turns, eye movements, or posture adjustments, using a fixed wide-angle view. Requirements: (1) customers browse or read in the foreground, shelves and several customers occupy the middle ground, and shelves extend into the background to show depth; (2) natural occlusion is clear, such as foreground people partially blocking background shelves, while relative positions remain stable; (3) a slightly crooked stack of books and a partly exposed handwritten recommendation card are clearly visible; (4) subject orientations and reading directions vary naturally rather than forming neat rows; (5) use documentary photography with a clear light direction, coordinated color across depth layers, and strong but non-rigid visual order; (6) motion is limited to stillness or very small natural changes, without obvious physical interaction or large page-turning actions. Use a fixed camera, natural daylight or soft illumination, and deep depth of field, emphasizing spatial depth, visual order, and photographic aesthetics in a complex composition. Primary rubric. (1) The bookstore contains at least 6 people, 4 shelving areas, and more than 30 visible books; all quantity requirements must be met. Additional criteria. (1) Foreground customers browse or read, the middle ground mixes several customers with shelves, and background shelves extend into the distance to create clear depth. (2) Foreground-background occlusion, light direction, and overall color treatment remain realistic and consistent. (3) The photographic treatment includes stable composition, natural daylight or soft illumination, and deep depth of field. (4) The visual treatment should convey spatial depth, visual order, and photographic aesthetics in a complex composition.	23. T2V_L2_D0_K0_0011 Prompt. A flower shop contains at least 4 people arranged around a central worktable as they organize bouquets. The table holds varied flowers, ribbons, and scissors, while people make only small posture adjustments. Use a fixed camera and emphasize the plausibility of a complex static scene. A fallen leaf beside the flower stems and small water droplets on a glass vase must be visible. Keep the distant background stable and avoid changes in local object counts. Multiple distinct subjects or objects must remain visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear depth layering, emphasizing layout, occlusion, and complex-composition stability. Primary rubric. (1) At least 4 people are positioned around the flower-shop worktable. (2) Flowers, ribbons, scissors, and the vase are clearly visible. Additional criteria. (1) The number and position of background flower racks remain stable. (2) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (3) The visual treatment should convey the plausibility of a complex static scene. (4) Multiple distinct subjects or objects remain visible with clear spatial relationships.
19. T2V_L2_D0_K0_0002 Prompt. An indoor market contains at least 8 customers, 3 stalls, and 12 groups of goods. People remain mostly still, making only very small observational or positional adjustments. Use a fixed camera with clear foreground, middle-ground, and background layers. A slightly reflective price sign and the edge of a hanging shopping bag must appear. Subject distribution should be balanced, with at least one natural occlusion between foreground and background. Motion is restricted to stillness or subtle observation, breathing, page turning, or stance adjustment, with no obvious physical interaction. Use a fixed wide-angle or mildly compressed telephoto documentary style, emphasizing layered depth, natural occlusion, balanced distribution, unified color, clear lighting direction, and visual order in a complex scene. Primary rubric. (1) The indoor market contains at least 8 customers, 3 stalls, and 12 groups of goods with balanced subject distribution. Additional criteria. (1) The market retains a documentary photographic style with unified color and a clear lighting direction. (2) The wide-angle view supports readable multi-subject spatial depth. (3) The visual treatment should convey layered depth, natural occlusion, balanced subject distribution, unified color, clear lighting direction, and visual order. (4) Supporting details include a slightly reflective price sign and the edge of a hanging shopping bag.	24. T2V_L2_D0_K0_0012 Prompt. In a bicycle-repair workshop, several people are distributed across different work stands in inspection or conversational poses. Tires, tools, and parts boxes are naturally arranged. Use a fixed camera and emphasize the realism of a static group scene. A horizontally placed wrench and a slightly tilted corner of a repair label must appear. Keep the distant background stable and avoid changes in local object counts. Multiple distinct subjects or objects must remain visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear depth layering, emphasizing layout, occlusion, and complex-composition stability. Primary rubric. (1) Several people are visible inside the bicycle-repair workshop. (2) Tires, tools, parts boxes, and work stands are clear. Additional criteria. (1) The number of background objects remains stable. (2) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (3) The visual treatment should convey the realism of a static group scene. (4) Supporting details include a horizontally placed wrench and a slightly tilted corner of a repair label.
20. T2V_L2_D0_K0_0004 Prompt. In an open-plan office, several people are distributed across different desks in working or conversational poses. Computers, papers, and cups are arranged naturally. Use a fixed camera and emphasize the realism of a static group scene. A horizontally placed pen and the slightly tilted corner of a sticky note must appear. Subject distribution should be balanced, with at least one natural foreground-background occlusion. Motion is restricted to stillness or subtle observation, breathing, page turning, or stance adjustment, without obvious physical interaction. Use a fixed wide-angle or mildly compressed telephoto documentary style, emphasizing layered depth, natural occlusion, balanced distribution, unified color, clear lighting direction, and visual order. Primary rubric. (1) A horizontally placed pen and the slightly tilted corner of a sticky note are clearly visible. Additional criteria. (1) Several people work or converse across the open office, while computers, papers, and cups are naturally distributed. (2) The office has realistic spatial layering, with unified color and a clear light direction. (3) The wide-angle view supports readable multi-subject spatial depth. (4) The visual treatment should convey the realism of a static group scene.	25. T2V_L2_D0_K0_0014 Prompt. A bus interior contains at least 8 passengers, 3 rows of seats, and 12 visible handrails or advertisements. People remain mostly still, making only small observational or positional adjustments. Use a fixed camera with clear foreground, middle-ground, and background layers. A slightly reflective transit card and the edge of a hanging backpack must appear. Standing and seated positions should remain clearly layered and the complex composition easy to read. Multiple distinct subjects or objects must remain visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear depth layering, emphasizing multi-subject layout, occlusion, and composition stability. Primary rubric. (1) At least 8 passengers are visible inside the bus. (2) Seats, handrails, advertisements, the transit card, and the backpack are visible. Additional criteria. (1) Standing and seated layers are clear. (2) Spatial relationships inside the bus remain stable. (3) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (4) The visual treatment should convey multi-subject layout, occlusion, and complex-composition stability.
21. T2V_L2_D0_K0_0005 Prompt. A train-station waiting hall contains at least 7 passengers, 3 rows of seats, multiple suitcases, and 2 information screens. People wait or look down at tickets, making only very small page turns or posture adjustments. Use a fixed camera and emphasize a natural multi-subject static composition. A wheeled suitcase with an old luggage tag and a partly exposed paper ticket must appear. Maintain reasonable spacing among the main people and keep the aisle or table edge readable. Multiple distinct subjects or objects must be visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear foreground-middle-background layering, emphasizing multi-subject layout, occlusion, and complex-composition stability. Primary rubric. (1) At least 7 passengers and multiple groups of waiting-hall seats are visible. (2) Suitcases, information screens, and the paper-ticket element are clear. Additional criteria. (1) Aisle edges and seat arrangement remain readable. (2) Foreground, middle-ground, and background layers remain stable. (3) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (4) The visual treatment should convey a natural multi-subject static composition.	26. T2V_L2_D0_K0_0020 Prompt. In a music rehearsal room, several people are distributed among different music stands in tuning or conversational poses. Instruments, sheet music, and water cups are naturally arranged. Use a fixed camera and emphasize the realism of a static group scene. A horizontally placed pencil and a slightly tilted metronome must appear. Subject orientations and attention points should vary naturally rather than forming a mechanical arrangement. Multiple distinct subjects or objects must remain visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear depth layering, emphasizing layout, occlusion, and complex-composition stability. Primary rubric. (1) Several people are visible inside the music rehearsal room. (2) Instruments, music stands, sheet music, and water cups are clearly distributed. Additional criteria. (1) Orientations and attention points are naturally varied. (2) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (3) The visual treatment should convey the realism of a static group scene. (4) Supporting details include a horizontally placed pencil and a slightly tilted metronome.
22. T2V_L2_D0_K0_0009 Prompt. A living room during a move contains at least 6 people, 4 groups of cardboard boxes, and multiple pieces of furniture. People remain in packing or inventory-checking poses, with only very small label-applying movements. Use a fixed camera and emphasize a natural multi-subject static composition. A slightly crooked stack of boxes and a partly exposed handwritten inventory list must appear. Keep the distant background stable and avoid changes in local object counts. Multiple distinct subjects or objects must be visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear foreground-middle-background layering, emphasizing layout, occlusion, and complex-composition stability. Primary rubric. (1) At least 6 people and multiple box areas are visible. (2) Furniture, boxes, and the handwritten inventory list are clear. Additional criteria. (1) The distant background remains stable. (2) The photographic treatment includes stable composition or a slow stable shot, relatively deep depth of field, and clear depth layering. (3) The visual treatment should convey a natural multi-subject static composition. (4) Supporting details include a slightly crooked stack of boxes and a partly exposed handwritten inventory list.	

Figure 7: Complete T2V benchmark cases for $S=2$, $D=0$, and $K=0$: high semantic complexity, standard detail, and no external knowledge requirement.

SHARED BASIC RUBRIC — applies to every task on this page (1) All static objects and background objects must remain visually consistent throughout the video. (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible. (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes. (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.	
27. T2V_L2_D0_K1_0001 Prompt. A laboratory tabletop contains at least 6 labeled containers, 5 pieces of glassware, 3 instruments, and 4 loose tools. At least four labels remain readable throughout the clip and must read "Acid A," "Base B," "pH 7 Buffer," and "Sample C." A pipette gun partly pinned beneath a record sheet and a small water mark at the bottom of a glass vessel must appear. All key labels, signs, and structural edges remain clear and stable throughout. Motion is limited to stillness or very small environmental changes, with no experimental reaction or obvious physical interaction. Use a fixed camera, professional laboratory or industrial documentary photography, even but directional lighting, deep depth of field, layered depth, and strict visual order, emphasizing object arrangement, label legibility, material detail, and overall aesthetics. Primary rubric. (1) At least 6 labeled containers, 5 pieces of glassware, 3 instruments, and 4 tools are visible. (2) At least four labels—"Acid A," "Base B," "pH 7 Buffer," and "Sample C"—are clearly readable. Additional criteria. (1) Object layout remains stable and clearly layered. (2) The complex laboratory-table composition is reasonable. (3) The visual treatment should convey professional object arrangement, label legibility, material detail, and overall aesthetics. (4) Supporting details include a pipette gun partly pinned beneath a record sheet and a small water mark at the bottom of a glass vessel.	31. T2V_L2_D0_K1_0006 Prompt. A radiology control console contains at least 4 image-preview screens, 5 examination forms, 3 controllers, and 4 loose marker notes. At least four text labels remain readable throughout and must read "Scan A," "Dose Log," "Patient ID," and "Series 03." A stylus partly pinned beneath a form and a small reflection point on a screen edge must appear. Spatial order, material consistency, text legibility, and compositional balance among the many objects should remain professional. Primary rubric. (1) The radiology console includes screens, forms, controllers, and marker notes. (2) "Scan A," "Dose Log," "Patient ID," and "Series 03" are readable. Additional criteria. (1) Spatial order and compositional balance remain professional. (2) Text legibility remains stable. (3) Supporting details include a stylus partly pinned beneath a form and a small reflection point on a screen edge. (4) Professional visual quality and clear visual order are maintained.
28. T2V_L2_D0_K1_0002 Prompt. An industrial workbench contains at least 2 control panels, 8 buttons or knobs, 3 warning signs, and 2 screens. Visible text must include "HIGH PRESSURE," "EMERGENCY STOP," and "Temp 68 C." A slightly coiled cable and a small maintenance label attached to a panel edge must appear. All key labels, signs, and structural edges remain clear and stable throughout. Motion is limited to stillness or very small environmental changes, with no experimental reaction or obvious physical interaction. Use a fixed camera, professional laboratory or industrial documentary photography, even but directional lighting, deep depth of field, layered depth, and strict visual order, emphasizing object arrangement, label legibility, material detail, and overall aesthetics. Primary rubric. (1) Two control panels, at least 8 buttons or knobs, 3 warning signs, and 2 screens are visible. (2) "HIGH PRESSURE," "EMERGENCY STOP," and "Temp 68 C" are clearly readable. (3) The coiled cable and maintenance label are present. (4) The panel layout remains stable. Additional criteria. (1) The visual treatment should convey professional object arrangement, label legibility, material detail, and overall aesthetics. (2) Supporting details include a slightly coiled cable and a small maintenance label attached to the panel edge. (3) Deep depth of field and layered depth strengthen visual order. (4) Professional laboratory or industrial documentary photography reinforces object arrangement and material detail.	32. T2V_L2_D0_K1_0007 Prompt. An aerospace-parts inspection bench contains at least 2 parts trays, 8 small fasteners or gauges, 3 warning signs, and 2 inspection drawings. Visible text must include "TORQUE LIMIT," "INSPECTED," and "Panel 68C." A slightly coiled antistatic wrist strap and a small maintenance label attached to a drawing edge must appear. Spatial order, material consistency, text legibility, and compositional balance among the many objects should remain professional. Primary rubric. (1) The inspection bench contains parts trays, fasteners, gauges, and drawings. (2) "TORQUE LIMIT," "INSPECTED," and "Panel 68C" are readable. Additional criteria. (1) Metallic and engineering materials remain consistent. (2) The complex static composition remains stable. (3) Supporting details include a slightly coiled antistatic wrist strap and a small maintenance label attached to a drawing edge. (4) Professional visual quality and clear visual order are maintained.
29. T2V_L2_D0_K1_0004 Prompt. A set of specimens or models with exhibit labels is neatly arranged on a display platform. At least 8 main objects and multiple signs are present, with sign text fixed as "Exhibit 07," "Gallery B," and "Specimen 12." The static scene emphasizes shape consistency and compositional logic under dense text labeling. Slight wear on one base edge and a partly obscured numbered card must appear. All key labels, signs, and structural edges remain clear and stable throughout. Motion is limited to stillness or very small environmental changes, with no experimental reaction or obvious physical interaction. Use a fixed camera, professional laboratory or industrial documentary photography, even but directional lighting, deep depth of field, layered depth, and strict visual order. Primary rubric. (1) At least 8 main objects and multiple signs are visible on the display platform. (2) "Exhibit 07," "Gallery B," and "Specimen 12" are displayed accurately. (3) The relationship between each sign and its corresponding exhibit is clear. Additional criteria. (1) The display composition is neat and stable. (2) The visual treatment should maintain shape consistency and compositional logic under dense text labeling. (3) Supporting details include slight wear on a base edge and a partly obscured numbered card. (4) Deep depth of field and layered depth strengthen visual order.	33. T2V_L2_D0_K1_0008 Prompt. A geological-sample cataloging table contains at least 4 rock trays, 5 sample boxes, 5 labeled objects, and 2 record sheets. Sampling has not begun, but the layout clearly suggests preparation for cataloging. Visible text must include "Basalt Core" and "Layer 1." A slight crease at the corner of a record sheet and a small amount of rock-powder residue beside a sample box must appear. Spatial order, material consistency, text legibility, and compositional balance should remain professional. Primary rubric. (1) Rock trays, sample boxes, labeled objects, and record sheets are visible. (2) "Basalt Core" and "Layer 1" are clear. Additional criteria. (1) The spatial arrangement of objects appears professional. (2) Relationships between text labels and samples are clear. (3) Supporting details include a slight crease on a record-sheet corner and rock-powder residue beside a sample box. (4) Professional visual quality and clear visual order are maintained.
30. T2V_L2_D0_K1_0005 Prompt. A professional experimental platform contains multiple numbered devices, measuring tools, and instruction labels in a complex composition. Visible text must include "Unit 04," "Step 2," and "Valve B." A horizontally placed marker and a thin wire draped over a tray edge must appear. Emphasize text legibility in a complex professional static scene and keep all key labels, signs, and structural edges clear and stable throughout. Motion is limited to stillness or very small environmental changes, with no experimental reaction or obvious physical interaction. Use a fixed camera, professional laboratory or industrial documentary photography, even but directional lighting, deep depth of field, layered depth, and strict visual order. Primary rubric. (1) Multiple devices, measuring tools, and instruction labels appear together. (2) "Unit 04," "Step 2," and "Valve B" are clearly readable. (3) The horizontal marker and thin wire draped over the tray edge are present. Additional criteria. (1) A professional experimental-platform appearance is maintained. (2) The visual treatment should emphasize text legibility in a complex professional static scene. (3) Supporting details include a horizontally placed marker and a thin wire draped over a tray edge. (4) Deep depth of field and layered depth strengthen visual order.	34. T2V_L2_D0_K1_0009 Prompt. At an electronics quality-inspection station, multiple circuit boards and test fixtures are neatly arranged on an antistatic mat. At least 8 main objects and multiple signs are present, with sign text fixed as "QC 07," "Board B," and "Reject 12." The static scene emphasizes shape consistency and compositional logic under dense text labeling. Slight wear on one fixture edge and a partly obscured numbered card must appear. Spatial order, material consistency, text legibility, and compositional balance should remain professional. Primary rubric. (1) The electronics inspection station contains circuit boards, test fixtures, and signs. (2) "QC 07," "Board B," and "Reject 12" are clear. Additional criteria. (1) The composition remains logical despite dense text. (2) The visual treatment should maintain shape consistency and compositional logic under dense text labeling. (3) Supporting details include slight wear on a fixture edge and a partly obscured numbered card. (4) Professional visual quality and clear visual order are maintained.

Figure 8: Complete T2V benchmark cases for $S=2$, $D=0$, and $K=1$: high semantic complexity, standard detail, and an external knowledge requirement.

SHARED BASIC RUBRIC — applies to every task on this page (1) All static objects and background objects must remain visually consistent throughout the video. (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible. (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes. (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.	
35. T2V_L2_D1_K0_0001 Prompt. A busy street simultaneously contains at least 10 pedestrians, 3 bicycles, 4 cars, and 5 storefront façades. Pedestrians, bicycles, and cars are all moving. Use a fixed wide-angle street view and emphasize complex dynamic relationships among many subjects. A slightly swaying roadside sign and a small bag placed beside a seat must appear. Motion directions in the foreground, middle ground, and background should all be clearly readable. Primary rubric. (1) At least 10 pedestrians, 3 bicycles, 4 cars, and 5 storefront façades are visible simultaneously. (2) Pedestrians, bicycles, and cars remain in continuous motion, with readable directions across all depth layers. (3) Different subject types maintain reasonable spatial distribution without abnormal overlap or disappearance. Additional criteria. (1) The wide-angle view supports readable multi-subject spatial depth. (2) The visual treatment should convey complex dynamic relationships among many subjects. (3) Supporting details include a slightly swaying roadside sign and a small bag beside a seat, while motion directions remain readable across depth layers.	39. T2V_L2_D1_K0_0005 Prompt. An airport terminal contains at least 9 travelers, 2 luggage carts, 6 groups of suitcases, and 4 counter or gate areas. Moving subjects appear in the foreground, middle ground, and background with natural rhythm. A half-open boarding bag and a slightly curled baggage tag at a counter edge must appear. At least one natural crossing or occlusion occurs among moving subjects. Multiple subjects form readable front-back and left-right relationships without random drift or identity confusion. Use relatively deep depth of field and a stable medium-long or slow tracking shot, emphasizing multi-subject dynamics, occlusion handling, and overall motion order. Primary rubric. (1) Travelers, luggage carts, suitcases, and counters are visible in the airport terminal. (2) Moving subjects appear in the foreground, middle ground, and background. (3) The half-open boarding bag and baggage tag are clear. Additional criteria. (1) Moving subjects exhibit natural crossing. (2) Spatial layering and subject identities remain stable. (3) The photographic treatment uses relatively deep depth of field and a stable medium-long or slow tracking shot. (4) The visual treatment should convey multi-subject dynamics, occlusion handling, and overall motion order.
36. T2V_L2_D1_K0_0002 Prompt. A market passage contains at least 9 pedestrians, 2 handcarts, 6 groups of boxes, and 4 stalls or storefronts. Moving subjects appear in the foreground, middle ground, and background with natural overall rhythm. A half-open cardboard box and a slightly curled paper sign at a stall edge must appear. Motion directions across all depth layers should be clearly readable. Primary rubric. (1) At least 9 pedestrians, 2 handcarts, 6 groups of boxes, and 4 stalls or storefronts are visible simultaneously. (2) Continuously moving subjects appear across all depth layers with clearly readable directions. (3) The half-open box and curled paper sign are clearly visible. Additional criteria. (1) Pedestrian and handcart motion is continuous and natural, preserving plausible traffic relationships. (2) The overall rhythm is stable and fully conveys the market's dynamic layering. (3) Supporting details include the half-open box and curled paper sign, while motion directions remain readable across depth layers. (4) The background atmosphere, supporting details, and visual rhythm remain consistent with the prompt.	40. T2V_L2_D1_K0_0006 Prompt. An indoor basketball gym contains at least 8 players, 2 basketballs, 4 training cones, and 1 courtside seating area. Players run, pass, or stand in different directions under a stable wide-angle camera. A strip of tape on the floor and a slightly swaying towel beside the seats must appear. At least one natural crossing or occlusion occurs among moving subjects. Multiple subjects form readable front-back and left-right relationships without random drift or identity confusion. Use relatively deep depth of field and a stable medium-long or slow tracking shot, emphasizing multi-subject dynamics, occlusion handling, and overall motion order. Primary rubric. (1) At least 8 players are moving inside the basketball gym. (2) Basketballs, training cones, and the courtside seating area are visible. (3) Running, passing, or defensive actions remain continuous. Additional criteria. (1) The tape mark and towel serve as supporting references. (2) Occlusion is natural and subject identities do not become confused. (3) The photographic treatment uses relatively deep depth of field and a stable medium-long or slow tracking shot. (4) The visual treatment should convey multi-subject dynamics, occlusion handling, and overall motion order.
37. T2V_L2_D1_K0_0003 Prompt. A playground contains at least 8 children, 1 ball, 4 playground structures, and 1 set of bleachers. The children run, pass the ball, and cross one another's paths. Use a fixed camera and emphasize a dynamic multi-subject scene. A water bottle at the edge of the field and a short chalk line on the ground must appear. Motion directions in the foreground, middle ground, and background should all be clearly readable. Primary rubric. (1) At least 8 children, 1 ball, 4 playground structures, and 1 bleacher area are visible simultaneously. (2) The children run, pass the ball, and cross paths, with the ball-transfer process continuously visible. (3) The water bottle and chalk line are clearly visible. (4) Moving subjects appear across all depth layers with clearly readable directions. Additional criteria. (1) Multi-subject interaction is continuous, natural, and consistent with playground activity. (2) The visual treatment should convey a dynamic multi-subject scene. (3) Supporting details include a water bottle and short chalk line, while motion directions remain readable across depth layers.	41. T2V_L2_D1_K0_0009 Prompt. An orchestral rehearsal hall contains at least 8 performers, multiple instruments, 4 music-stand areas, and 1 observation-seating area. Following the conductor's cues, performers make subtle but synchronized body and arm movements. Use a stable wide-angle camera. A pencil beside a music stand and one slightly fluttering sheet of music in the front seating row must appear. Main motion paths remain continuous without abrupt jumps or broken rhythm. Multiple subjects form readable front-back and left-right relationships without random drift or identity confusion. Use relatively deep depth of field and a stable medium-long or slow tracking shot, emphasizing multi-subject dynamics, occlusion handling, and overall motion order. Primary rubric. (1) At least 8 performers and multiple instruments are visible. (2) Subtle synchronized movement occurs under the conductor's cues. (3) The music stands, pencil, and slightly fluttering sheet music are clear. Additional criteria. (1) The wide-angle composition remains stable. (2) Group identities and spatial relationships do not become confused. (3) The photographic treatment uses relatively deep depth of field and a stable medium-long or slow tracking shot. (4) The visual treatment should convey multi-subject dynamics, occlusion handling, and overall motion order.
38. T2V_L2_D1_K0_0004 Prompt. A port loading area simultaneously contains at least 10 workers, 2 forklifts, 3 handcarts, and the edges of 5 container groups. Workers, forklifts, and handcarts are moving, with at least one natural crossing or occlusion among moving subjects. Multiple subjects must form readable front-back and left-right relationships without random drift or identity confusion. Use relatively deep depth of field and a stable medium-long shot or slow tracking shot, emphasizing multi-subject dynamics, occlusion handling, and overall motion order. Primary rubric. (1) Workers, forklifts, handcarts, and containers are visible in the port loading area. (2) At least 10 workers and multiple moving vehicles are visible. (3) Moving subjects exhibit natural crossing or occlusion. Additional criteria. (1) Front-back and left-right relationships remain clear. (2) The photographic treatment uses relatively deep depth of field and a stable medium-long or slow tracking shot. (3) The visual treatment should convey multi-subject dynamics, occlusion handling, and overall motion order. (4) Deep depth of field and layered depth strengthen visual order.	42. T2V_L2_D1_K0_0010 Prompt. A warehouse sorting area simultaneously contains at least 10 workers, 3 handcarts, 4 conveyor segments, and 5 shelving façades. Workers, handcarts, and parcels are moving, while the temporal order and path relationships among moving subjects remain clear. Multiple subjects form readable front-back and left-right relationships without random drift or identity confusion. Use relatively deep depth of field and a stable medium-long or slow tracking shot, emphasizing multi-subject dynamics, occlusion handling, and overall motion order. Primary rubric. (1) Workers, handcarts, conveyors, and shelves are visible in the warehouse sorting area. (2) At least 10 workers participate in the motion. (3) Path relationships among parcels and people are clear. Additional criteria. (1) The temporal order remains traceable. (2) Overall motion order remains stable. (3) The photographic treatment uses relatively deep depth of field and a stable medium-long or slow tracking shot. (4) The visual treatment should convey multi-subject dynamics, occlusion handling, and overall motion order.

Figure 9: Complete T2V benchmark cases for $S=2$, $D=1$, and $K=0$: high semantic complexity, rich detail, and no external knowledge requirement.

SHARED BASIC RUBRIC — applies to every task on this page

- (1) All static objects and background objects must remain visually consistent throughout the video.
- (2) The video should appear realistic. In particular, when material reactions or object interactions are involved, and unless the prompt states otherwise, dynamic objects should follow real-world physical, material, and chemical laws and exhibit logically consistent changes. Material quantities should be conserved as strictly as possible.
- (3) For videos containing motion, each action should be complete and reasonable. Unless otherwise specified, all actions and reactions should remain continuous and natural, without obvious abrupt changes.
- (4) If a task-specific instruction conflicts with the three requirements above, such as a scene that intentionally disregards physical laws, the task-specific instruction takes precedence; the three shared requirements are secondary.

43. T2V_L2_D1_K1_0001

Prompt.

A professional wide-angle recording of a 100-meter sprint contains at least 8 athletes, a clearly visible track, and a timing environment. The race is shown continuously from start to finish, with athletes' relative positions gradually changing as the event progresses. The composition is complex and professional. A small towel beside a starting block and a slightly swaying handheld sign in the front row of the stands must appear. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

Primary rubric.

- (1) The start, acceleration, mid-race running, and finish stages appear in sequence.
- (2) The small towel beside the starting block remains identifiable.
- (3) The handheld sign in the front row of the stands sways slightly.
- (4) The changing positions of multiple athletes remain continuously traceable.

Additional criteria.

- (1) The wide-angle view supports readable multi-subject spatial depth.
- (2) Supporting details include the small towel beside the starting block and the slightly swaying handheld sign.
- (3) Professional visual quality and clear visual order are maintained.

44. T2V_L2_D1_K1_0002

Prompt.

A multi-person dance performance on a professional stage contains at least 8 dancers. The dancers complete a continuous unified choreography with observable changes in group positions. The camera preserves the full stage space and group interaction. A floor light at the side of the stage remains partly visible, and a fabric accessory at one dancer's waist sways slightly. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

Primary rubric.

- (1) At least 8 dancers participate simultaneously.
- (2) Formation changes are clear and complete.
- (3) The floor light at the side of the stage remains visible.
- (4) Group movement is synchronized and spatial relationships remain stable.

Additional criteria.

- (1) The specified dancer's waist accessory sways naturally.
- (2) Supporting details include the partly visible floor light and the slightly swaying fabric accessory.
- (3) Professional visual quality and clear visual order are maintained.

45. T2V_L2_D1_K1_0003

Prompt.

In a complex multi-object mechanics experiment, multiple devices, measuring tools, and target objects are visible together. The apparatus is triggered in a specified manner, after which the objects continue interacting and produce subsequent changes. A measurement scale sheet attached to a support and a spare weight block at a corner of the workbench must appear. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

Primary rubric.

- (1) Multiple devices, measuring tools, and target objects are visible together.
- (2) The trigger condition, main phenomenon, and final state are clearly presented.
- (3) The measurement scale sheet on the support remains visible.
- (4) The spare weight block remains at the corner of the workbench.
- (5) Interactions among the devices remain continuously traceable.

Additional criteria.

- (1) Supporting details include the measurement scale sheet on the support and the spare weight block at the workbench corner.
- (2) The background atmosphere, supporting details, and visual rhythm remain consistent with the prompt.

46. T2V_L2_D1_K1_0004

Prompt.

In an aerial formation or complex stunt scene, multiple high-speed subjects execute continuous flight maneuvers simultaneously. Their relative positions change continuously and gradually form new spatial relationships. A small flag in the ground observation area and a faint, short-lived smoke trail behind one aircraft must appear. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

Primary rubric.

- (1) Multiple flying subjects complete a formation change.
- (2) Motion stages and relative-position changes are clearly presented.
- (3) The small flag in the ground observation area remains visible.
- (4) At least one aircraft leaves a brief faint smoke trail.
- (5) The flying subjects gradually form a new spatial relationship.

Additional criteria.

- (1) Supporting details include the small flag in the observation area and the faint smoke trail behind one aircraft.
- (2) The background atmosphere, supporting details, and visual rhythm remain consistent with the prompt.

47. T2V_L2_D1_K1_0005

Prompt.

On a complex chemical experimental platform, materials from multiple containers are added to a main container in a specified sequence. The states of the materials, addition order, and operation process remain clearly visible, and subsequent changes follow common real-world physical or chemical laws. A record card marked by a small liquid droplet and a small amount of residue on the end of a stirring rod must appear. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

Primary rubric.

- (1) Multiple materials are added or mixed sequentially.
- (2) The mixing process remains continuously visible.
- (3) An observable change appears inside the container.
- (4) The change occurs after mixing begins.
- (5) The full process has a clear temporal order.

Additional criteria.

- (1) Supporting details include the droplet-marked record card and residue on the end of the stirring rod.
- (2) The droplet and stirring-rod residue provide experimental reference details.

48. T2V_L2_D1_K1_0010

Prompt.

On a robotic assembly line, multiple robotic arms sequentially grasp, rotate, and place parts in a specified order. Part states, transfer order, and the operating process remain clearly visible, and subsequent changes follow common real-world mechanical-assembly logic. A calibration card attached to the workstation edge and a small amount of lubricant residue on a robotic-arm end effector must appear. Relative-position changes among people or devices remain fully traceable, while the scene preserves a stable, credible experimental environment and professional operating procedure.

Primary rubric.

- (1) The scene is a robotic assembly line.
- (2) Multiple robotic arms grasp, rotate, and place parts in sequence.
- (3) The transfer order and relative-position changes remain fully traceable.

Additional criteria.

- (1) The calibration card and lubricant residue serve as supporting references.
- (2) The mechanical-assembly logic remains stable and credible.
- (3) Supporting details include the calibration card and lubricant residue; relative-position changes remain fully traceable in a stable, credible environment and professional workflow.
- (4) Professional visual quality and clear visual order are maintained.

Figure 10: Complete T2V benchmark cases for $S=2$, $D=1$, and $K=1$: high semantic complexity, rich detail, and an external knowledge requirement.

T2V Case Example

Moderate semantics · Rich detail · No external knowledge · six model generations

S1-D1-K0

PROMPT

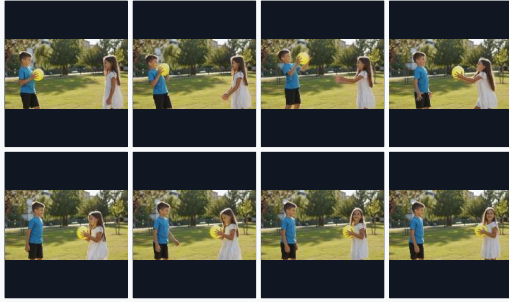
One child throws a yellow ball to another child, who catches it and steps slightly backward. Use a fixed camera. Emphasize direct interaction between a small number of subjects, preserve natural continuity across the throw and catch, and make the action order and rhythm of both participants clear.

TASK-SPECIFIC PRIMARY RUBRIC

1. The ball travels toward the second child along a continuous trajectory.
2. The second child successfully catches the ball and reacts by stepping slightly backward.
3. The full sequence of throwing, flight, and catching is complete and continuous.
4. The spatial relationship between the two children and the outcome of the interaction remain clearly visible.

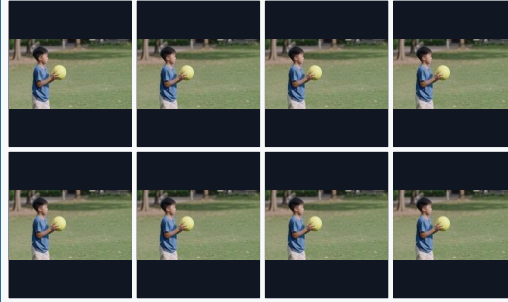
Seedance 2

6.0 s



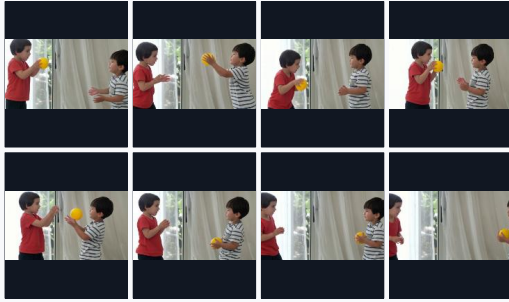
Kling 3.0 Pro

6.0 s



WAN 2.7

6.0 s



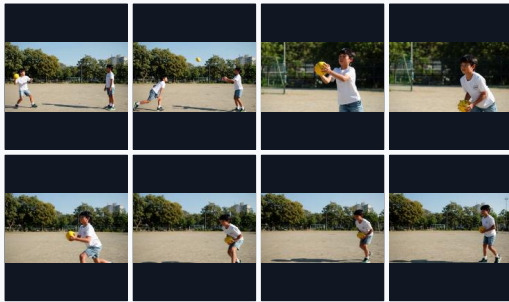
Veo 3.1 Fast

6.0 s



PixVerse C1

6.0 s



LTX 2.3 Pro

6.1 s



RAVEN-Eval supplementary material · T2V · S1-D1-K0 · frames sampled at fixed relative positions · 1/7

Figure 11: A representative T2V case under the S1-D1-K0 configuration. The case evaluates a continuous ball-throwing and catching interaction between two children without requiring external knowledge.

T2V Case Example

Moderate semantics · Rich detail · External knowledge · six model generations

S1-D1-K1

PROMPT

One astronaut slowly walks toward a giant black monolith while a second astronaut stands beside a rover and observes. Use a 2001: A Space Odyssey-style austere science-fiction aesthetic with symmetrical composition, minimalist space, slow movement, cool white light, and extensive silent negative space. Motion should be subtle but continuous, emphasizing cinematic composition and a solemn atmosphere.

TASK-SPECIFIC PRIMARY RUBRIC

1. The frame contains two astronauts and one giant black monolith.
2. One astronaut slowly approaches the monolith while the other observes beside the rover.
3. The scene presents an austere, minimalist science-fiction aesthetic in the style of 2001: A Space Odyssey.

Seedance 2 Fast

6.0 s



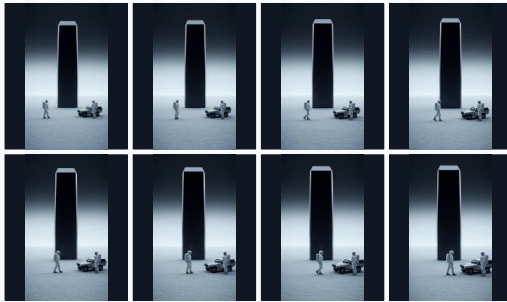
Kling 3.0 Omni

6.0 s



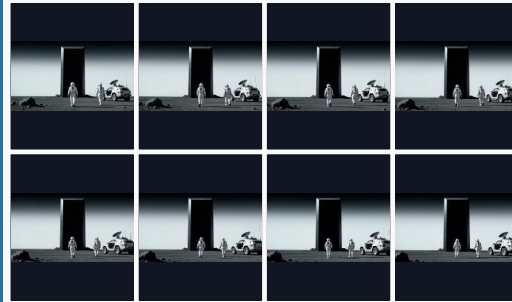
Grok-Image-Video

6.0 s



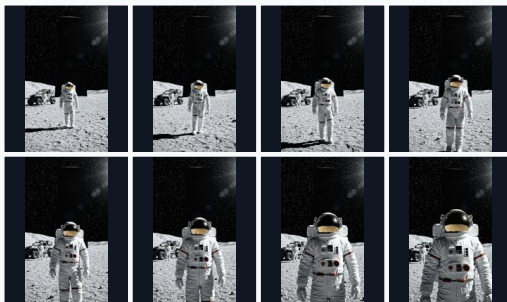
WAN 2.6

6.0 s



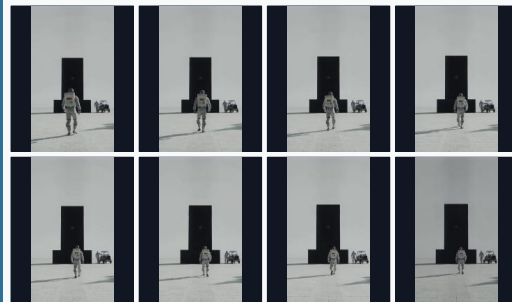
Veo 3.1 Lite

6.0 s



LTX 2.3 Fast

6.1 s



RAVEN-Eval supplementary material · T2V · S1-D1-K1 · frames sampled at fixed relative positions · 2/7

Figure 12: A representative T2V case under the S1-D1-K1 configuration. The case requires a restrained science-fiction composition involving two astronauts and a monolith while following the specified cinematic conventions.

T2V Case Example

High semantics · Standard detail · No external knowledge · six model generations

S2-D0-K0

PROMPT

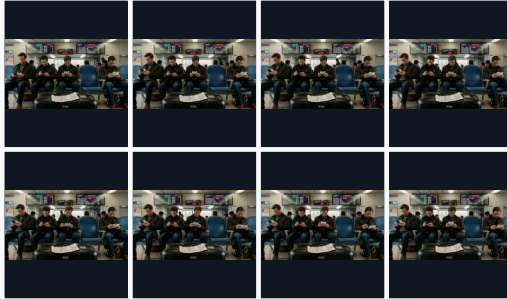
A train-station waiting hall contains at least 7 passengers, 3 rows of seats, multiple suitcases, and 2 information screens. People wait or look down at tickets, making only very small page turns or posture adjustments. Use a fixed camera and emphasize a natural multi-subject static composition. A wheeled suitcase with an old luggage tag and a partly exposed paper ticket must appear. Maintain reasonable spacing among the main people and keep the aisle or table edge readable. Multiple distinct subjects or objects must be visible with clear spatial relationships. Use a fixed camera or slow stable shot, relatively deep depth of field, and clear foreground-middle-background layering, emphasizing multi-subject layout, occlusion, and complex-composition stability.

TASK-SPECIFIC PRIMARY RUBRIC

1. At least 7 passengers and multiple groups of waiting-hall seats are visible.
2. Suitcases, information screens, and the paper-ticket element are clear.

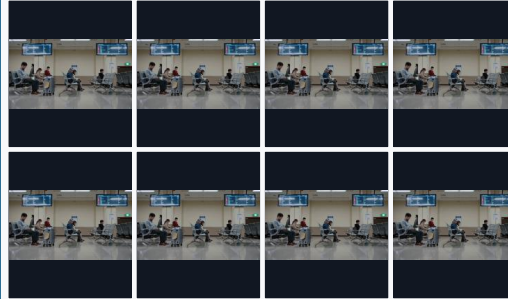
Seedance 1.5 Pro

6.0 s



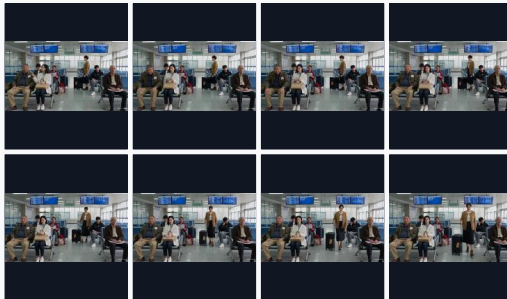
Kling 3.0 Std

6.0 s



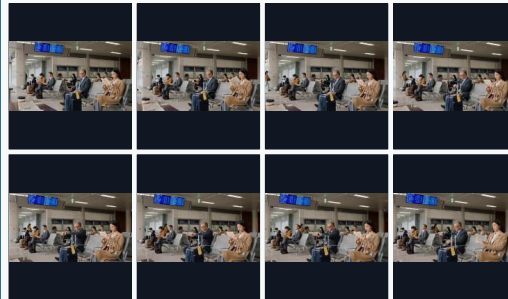
HappyHorse V1.0

6.1 s



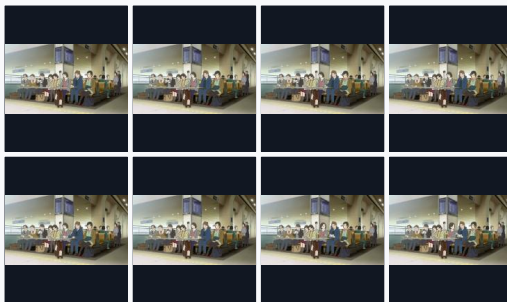
PixVerse V6

6.0 s



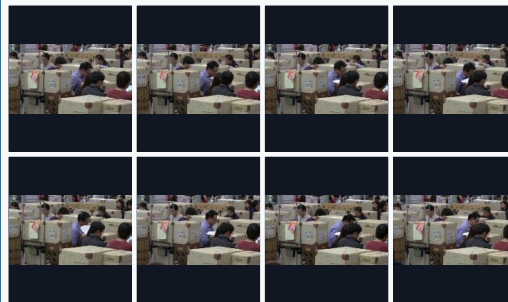
Hailuo 2.3

5.9 s



LTX 2.0 Pro

6.1 s



RAVEN-Eval supplementary material · T2V · S2-D0-K0 · frames sampled at fixed relative positions · 3/7

Figure 13: A representative T2V case under the S2-D0-K0 configuration. The case evaluates the stability of a crowded train-station waiting hall containing multiple passengers, seats, suitcases, and information screens.

T2V Case Example

High semantics · Standard detail · External knowledge · six model generations

S2-D0-K1

PROMPT

A laboratory tabletop contains at least 6 labeled containers, 5 pieces of glassware, 3 instruments, and 4 loose tools. At least four labels remain readable throughout the clip and must read "Acid A," "Base B," "pH 7 Buffer," and "Sample C." A pipette gun partly pinned beneath a record sheet and a small water mark at the bottom of a glass vessel must appear. All key labels, signs, and structural edges remain clear and stable throughout. Motion is limited to stillness or very small environmental changes, with no experimental reaction or obvious physical interaction. Use a fixed camera, professional laboratory or industrial documentary photography, even but directional lighting, deep depth of field, layered depth, and strict visual order, emphasizing object arrangement, label legibility, material detail, and overall aesthetics.

TASK-SPECIFIC PRIMARY RUBRIC

1. At least 6 labeled containers, 5 pieces of glassware, 3 instruments, and 4 tools are visible.
2. At least four labels—"Acid A," "Base B," "pH 7 Buffer," and "Sample C"—are clearly readable.

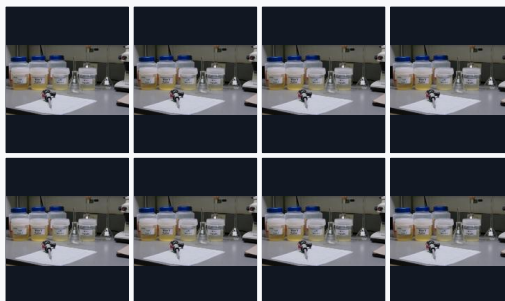
Seedance 2

6.0 s



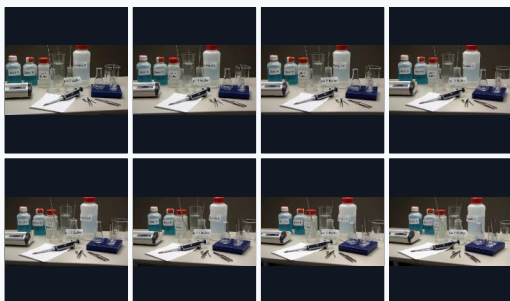
Kling 3.0 Pro

6.0 s



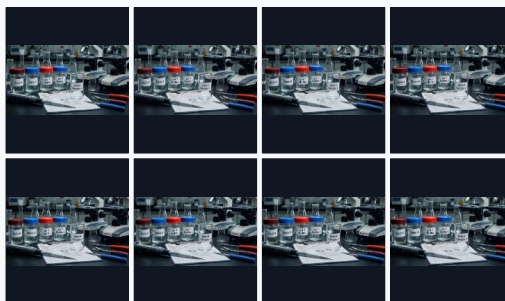
WAN 2.7

6.0 s



PixVerse V5.5

5.0 s



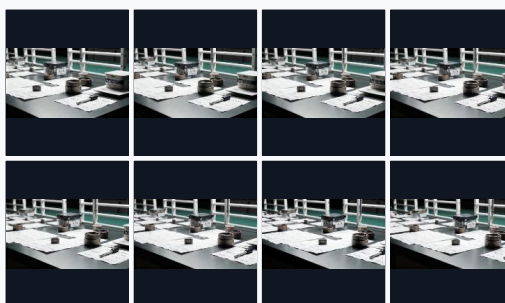
Veo 3.1 Lite

6.0 s



LTX 2.0 Fast

6.1 s



RAVEN-Eval supplementary material · T2V · S2-D0-K1 · frames sampled at fixed relative positions · 4/7

Figure 14: A representative T2V case under the S2-D0-K1 configuration. The case evaluates object-count compliance, text legibility, material consistency, and spatial organization on a densely populated laboratory tabletop.

T2V Case Example

High semantics · Rich detail · No external knowledge · six model generations

S2-D1-K0

PROMPT

A playground contains at least 8 children, 1 ball, 4 playground structures, and 1 set of bleachers. The children run, pass the ball, and cross one another's paths. Use a fixed camera and emphasize a dynamic multi-subject scene. A water bottle at the edge of the field and a short chalk line on the ground must appear. Motion directions in the foreground, middle ground, and background should all be clearly readable.

TASK-SPECIFIC PRIMARY RUBRIC

1. At least 8 children, 1 ball, 4 playground structures, and 1 bleacher area are visible simultaneously.
2. The children run, pass the ball, and cross paths, with the ball-transfer process continuously visible.
3. The water bottle and chalk line are clearly visible.
4. Moving subjects appear across all depth layers with clearly readable directions.

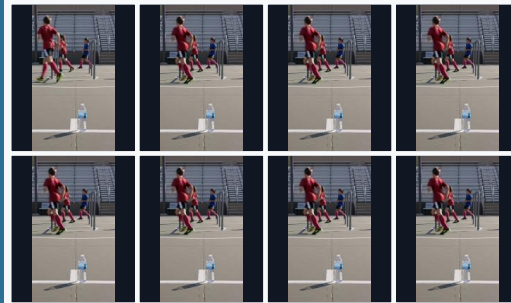
Seedance 2 Fast

6.0 s



Kling 3.0 Omni

6.0 s



Grok-Image-Video

6.0 s



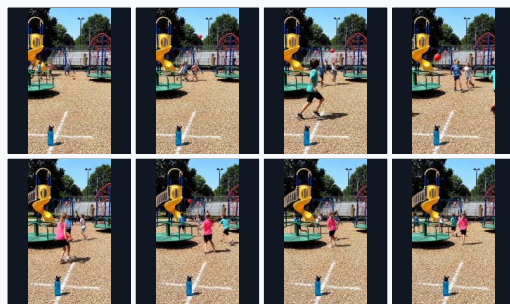
WAN 2.6

6.0 s



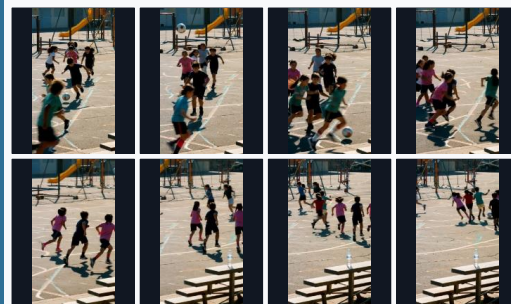
Veo 3.1 Fast

6.0 s



PixVerse V6

6.0 s



RAVEN-Eval supplementary material · T2V · S2-D1-K0 · frames sampled at fixed relative positions · 5/7

Figure 15: A representative T2V case under the S2-D1-K0 configuration. The case evaluates multi-subject motion, ball passing, path crossings, and readable movement across foreground, middle-ground, and background regions.

T2V Case Example

High semantics · Rich detail · External knowledge · six model generations

S2-D1-K1

PROMPT

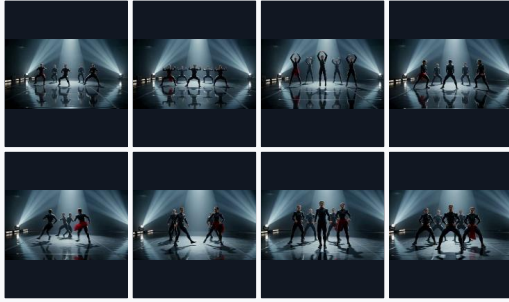
A multi-person dance performance on a professional stage contains at least 8 dancers. The dancers complete a continuous unified choreography with observable changes in group positions. The camera preserves the full stage space and group interaction. A floor light at the side of the stage remains partly visible, and a fabric accessory at one dancer's waist sways slightly. Transitions between key stages remain clear and supporting reference objects stay visible throughout the complex motion.

TASK-SPECIFIC PRIMARY RUBRIC

1. At least 8 dancers participate simultaneously.
2. Formation changes are clear and complete.
3. The floor light at the side of the stage remains visible.
4. Group movement is synchronized and spatial relationships remain stable.

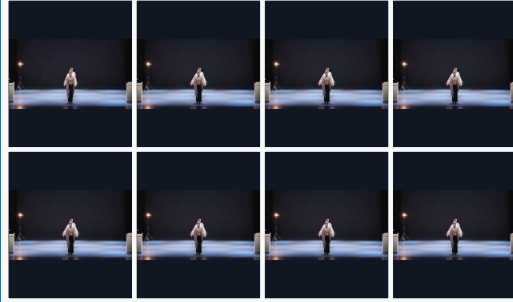
Seedance 1.5 Pro

6.0 s



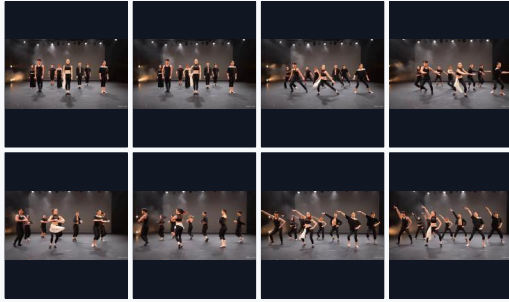
Kling 3.0 Std

6.0 s



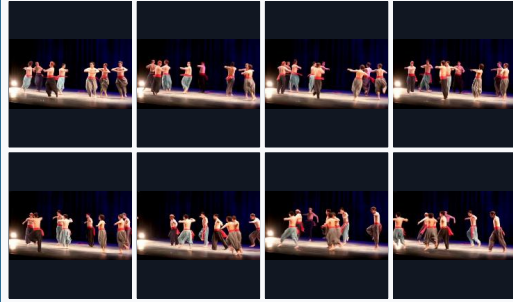
HappyHorse V1.0

6.1 s



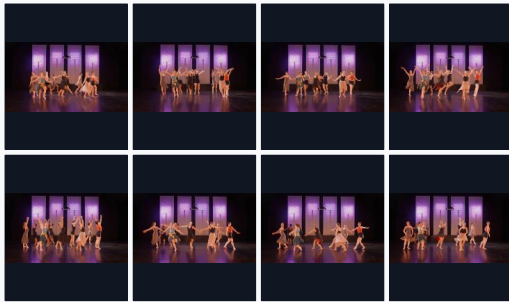
WAN 2.7

6.0 s



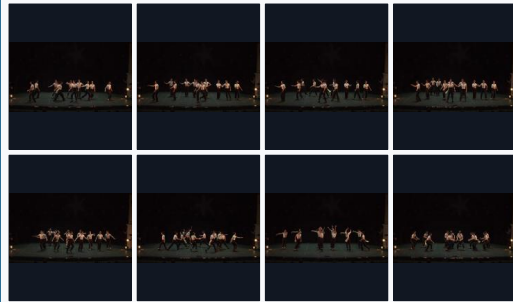
Hailuo 2.3

5.9 s



LTX 2.3 Pro

6.1 s



RAVEN-Eval supplementary material · T2V · S2-D1-K1 · frames sampled at fixed relative positions · 6/7

Figure 16: A representative T2V case under the S2-D1-K1 configuration. The case requires at least eight dancers to perform synchronized choreography with coherent formation changes and stable spatial relationships.

T2V Case Example

Physical outcome inference · Material response · six model generations

S1-D1-K1

Reasoning

PROMPT

A hydraulic press slowly descends toward a rubber duck, with the press head, duck, and workbench visible together. The complete process should show the press head moving continuously downward, making contact with the target, and producing subsequent changes that follow the natural response of a material under force. The interaction path, final state, and professional scene logic should remain clear and explainable.

TASK-SPECIFIC PRIMARY RUBRIC

1. The press head continuously approaches the target object.
2. The press head makes clear contact with the target.
3. The target visibly deforms after force is applied.
4. The final state is consistent with the applied-force process.

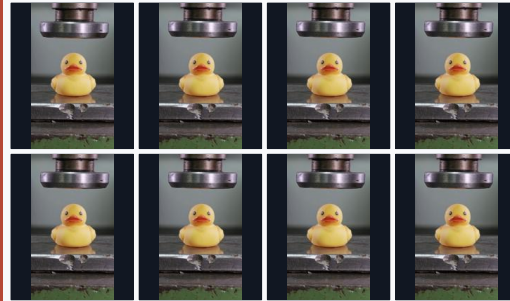
Seedance 2

6.0 s



Kling 3.0 Pro

6.0 s



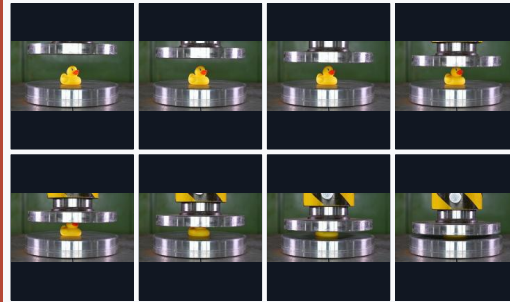
Grok-Image-Video

6.0 s



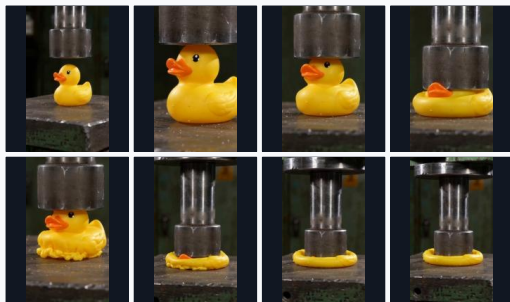
WAN 2.6

6.0 s



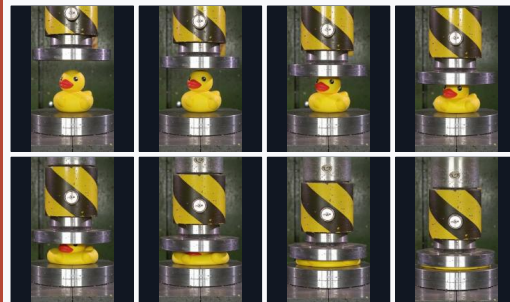
PixVerse C1

6.0 s



LTX 2.3 Fast

6.1 s



RAVEN-Eval supplementary material · T2V · Reasoning · frames sampled at fixed relative positions · 7/7

Figure 17: A representative reasoning-focused T2V case. The models must infer and generate the material response of a rubber duck as a hydraulic press descends, makes contact, and applies force.

FLT Case Example

Reference sequence and six model generations

Reasoning

YouTube Source

PROMPT

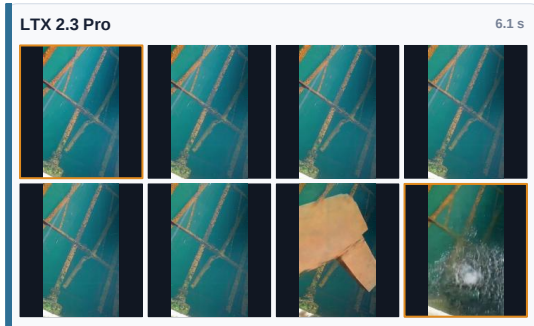
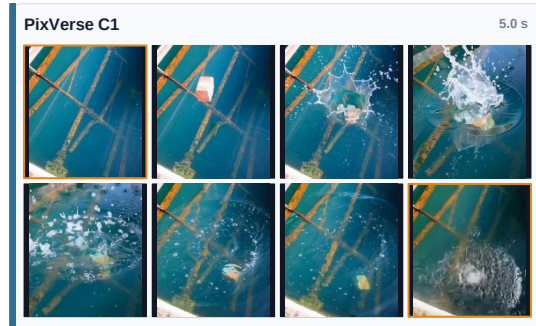
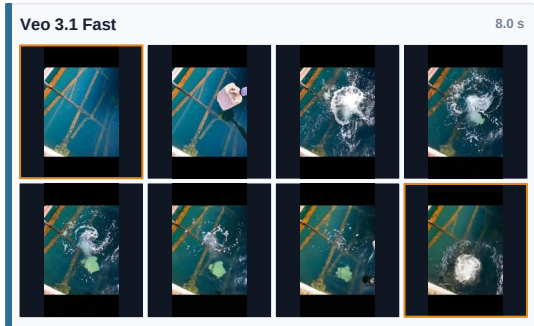
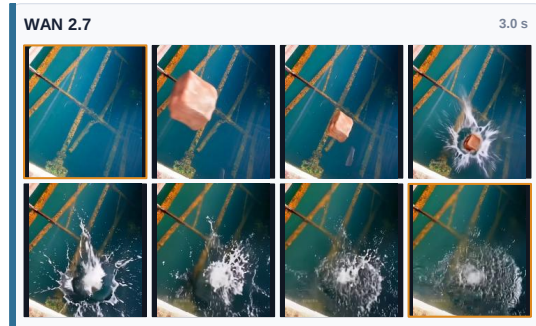
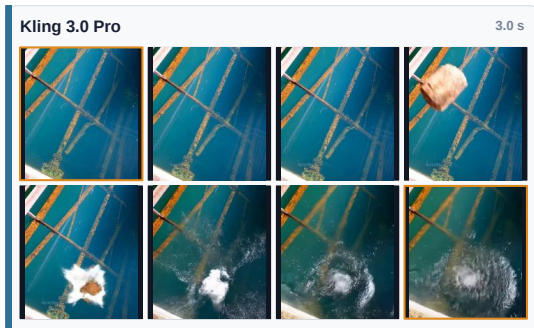
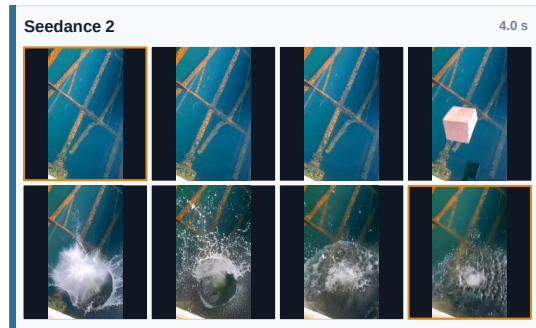
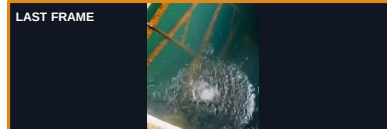
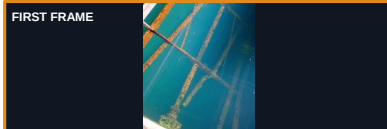
A large brick is suddenly thrown into the still surface of a sewage pool.
Complete the entire process.

TASK-SPECIFIC PRIMARY RUBRIC

1. A large brick must be thrown into the initially still sewage pool.
2. A large splash should appear immediately, followed by several clearly visible concentric ripples.

MODEL INPUT · JSON-defined first and last frames

JSON t = 12.706 / 15.581 s



RAVEN-Eval supplementary material · FLT · Reasoning · YouTube · generated frames: first + 6 uniformly spaced intermediate + last · 1/8

Figure 18: A reasoning FLT case collected from YouTube. The first and last frames are provided as model inputs. Six model outputs are shown using eight uniformly sampled frames per model.

FLT Case Example

Reference sequence and six model generations

Reasoning

ASMR Source

PROMPT

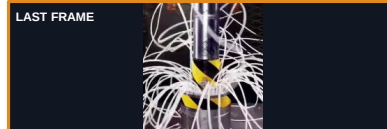
Show the complete process of a hydraulic press compressing several white steamed buns: the buns first flatten, and then thin white strips are extruded through holes in the press plate.

TASK-SPECIFIC PRIMARY RUBRIC

1. The full sequence must show the press moving downward, the buns flattening, and thin white strips being forced through the holes in the press plate.

MODEL INPUT · JSON-defined first and last frames

JSON t = 25.452 / 28.444 s



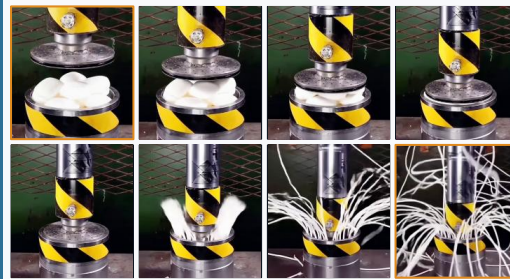
Seedance 2 Fast

4.0 s



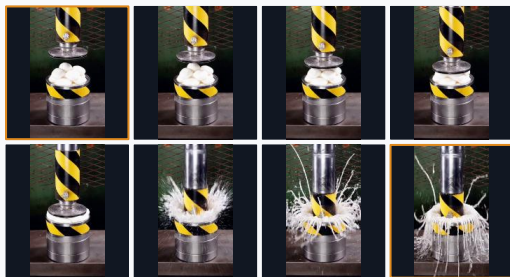
Kling 3.0 Omni

3.0 s



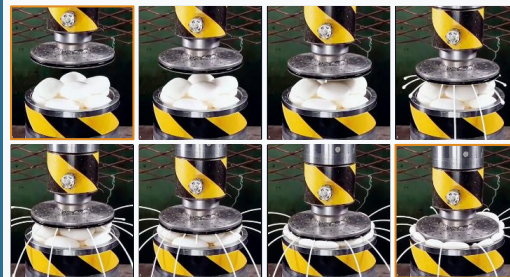
Grok-Image-Video

3.0 s



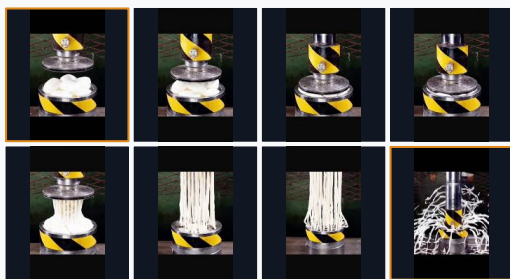
WAN 2.6

4.0 s



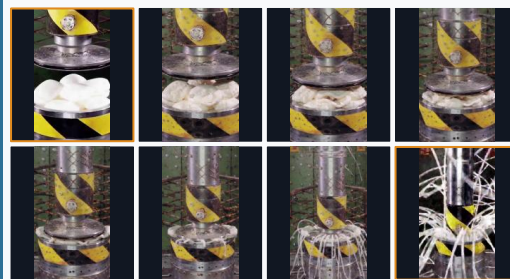
Veo 3.1 Lite

8.0 s



LTX 2.3 Fast

6.1 s



RAVEN-Eval supplementary material · FLT · Reasoning · ASMR · generated frames: first + 6 uniformly spaced intermediate + last · 2/8

Figure 19: A reasoning FLT case collected from an ASMR video. The case requires the models to infer and generate the physical transition between the given first and last frames.

FLT Case Example

Reference sequence and six model generations

Non-reasoning

YouTube Source

PROMPT

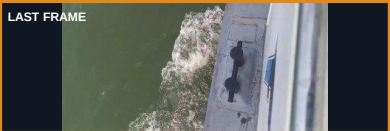
Keep the top-down viewpoint identical to the first frame, steady and free of motion blur. As the boat moves, white wake continuously forms beneath it and spreads outward with clearly discernible texture.

TASK-SPECIFIC PRIMARY RUBRIC

- 1. The camera must remain in the same stable top-down view as the first frame, without motion blur.
- 2. The boat's motion must continuously generate a spreading wake with clear white-water texture.

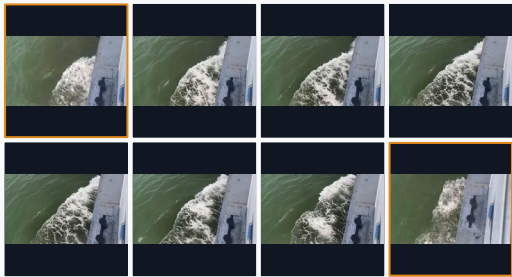
MODEL INPUT · JSON-defined first and last frames

JSON t = 59.570 / 61.990 s



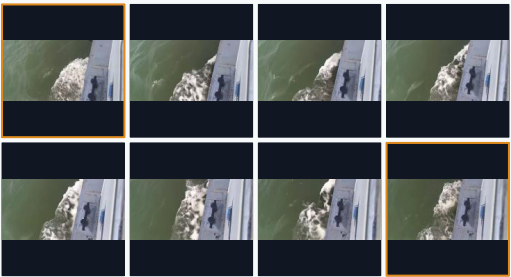
Seedance 1.5 Pro

4.0 s



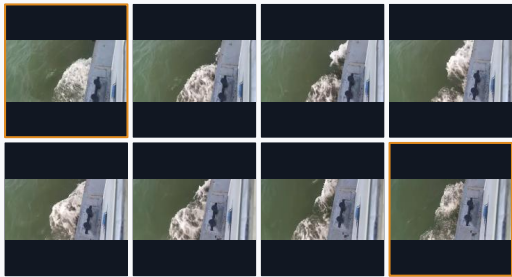
Kling 3.0 Std

3.0 s



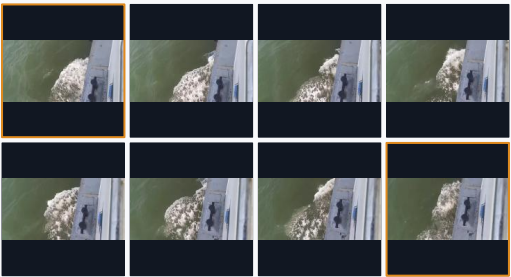
WAN 2.7

2.0 s



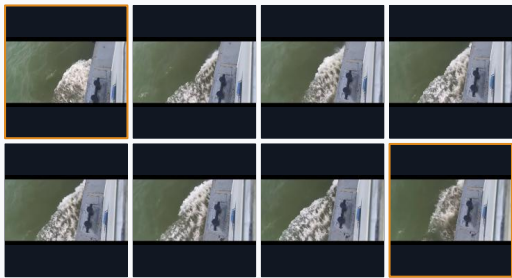
PixVerse V6

5.0 s



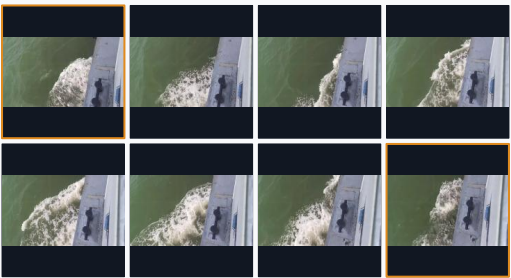
Veo 3.1 Fast

8.0 s



LTX 2.3 Pro

6.1 s



RAVEN-Eval supplementary material · FLT · Non-reasoning · YouTube · generated frames: first + 6 uniformly spaced intermediate + last · 3/8

Figure 20: A non-reasoning FLT case collected from YouTube. The case primarily evaluates visual continuity, camera consistency, and the temporal development of the boat wake.

FLT Case Example

Reference sequence and six model generations

Non-reasoning

ASMR Source

PROMPT

Show two hands kneading the petal-shaped soap into fragments, including the detailed breakage caused by the applied hand pressure.

TASK-SPECIFIC PRIMARY RUBRIC

1. The two hands must knead the petal-shaped soap into fragments, and the crushing details produced by the applied force must be visible.

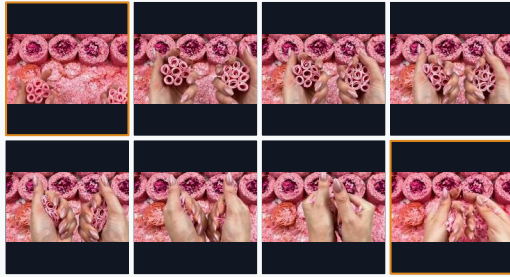
MODEL INPUT · JSON-defined first and last frames

JSON t = 281.890 / 285.324 s



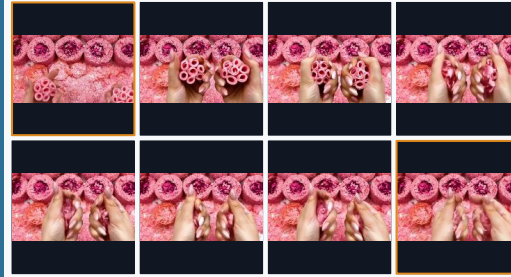
Seedance 2

4.0 s



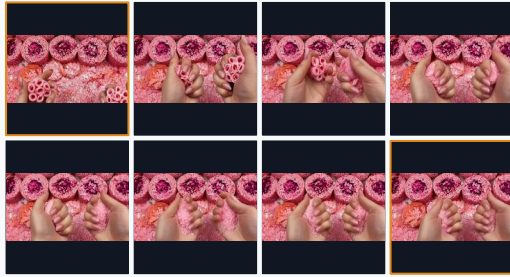
Kling 3.0 Pro

3.0 s



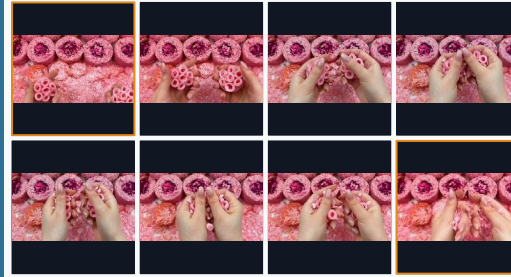
WAN 2.6

4.0 s



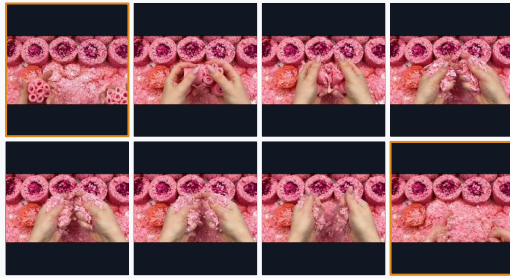
PixVerse V5.5

5.0 s



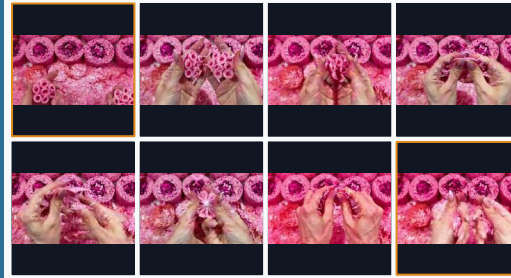
Veo 3.1 Lite

8.0 s



LTX 2.3 Fast

6.1 s



RAVEN-Eval supplementary material · FLT · Non-reasoning · ASMR · generated frames: first + 6 uniformly spaced intermediate + last · 4/8

Figure 21: A non-reasoning FLT case collected from an ASMR video. The case evaluates whether the models faithfully complete the soap-crushing process between the input frames.

KFT Case Example

Reference sequence and six model generations

Reasoning

YouTube Source

PROMPT

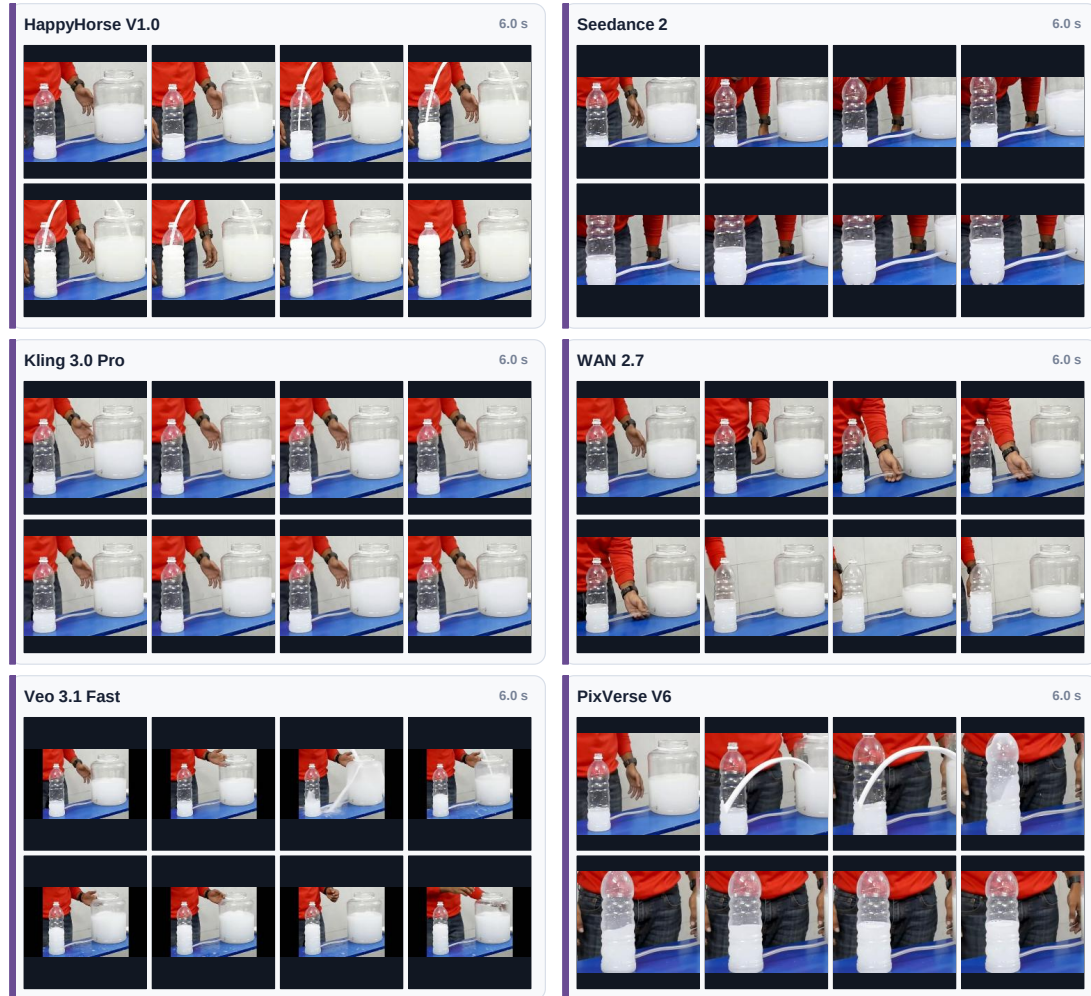
Using the communicating-vessels principle, infer the final equilibrium water levels in the large bucket and the small bottle, and generate the complete water-level transition. Keep the liquid color and container shapes consistent throughout.

TASK-SPECIFIC PRIMARY RUBRIC

1. The water level in the small bottle should rise slowly until it visually matches the level in the large bucket.
2. The liquid color and the shapes of both containers must remain consistent throughout.

MODEL INPUT · starting keyframe

0.7 s reference



RAVEN-Eval supplementary material · KFT · Reasoning · YouTube · frames sampled at fixed relative positions · 5/8

Figure 22: A reasoning KFT case collected from YouTube. Starting from the given keyframe, the models must infer the water-level transition according to the communicating-vessels principle.

KFT Case Example

Reference sequence and six model generations

Reasoning

ASMR Source

PROMPT

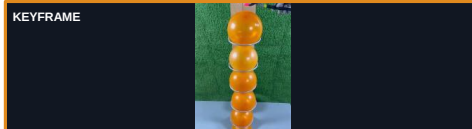
A red-hot iron ball falls freely from above six balloons. Predict and show the complete subsequent phenomenon.

TASK-SPECIFIC PRIMARY RUBRIC

1. Because of its high temperature, the falling iron ball should burst all six balloons from top to bottom almost instantaneously.
2. The six bursts should be extremely fast but still continuous, with their order discernible.

MODEL INPUT · starting keyframe

0.7 s reference



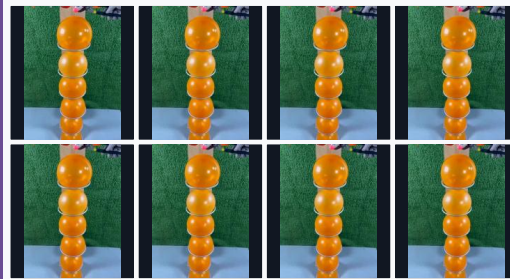
Seedance 2 Fast

6.0 s



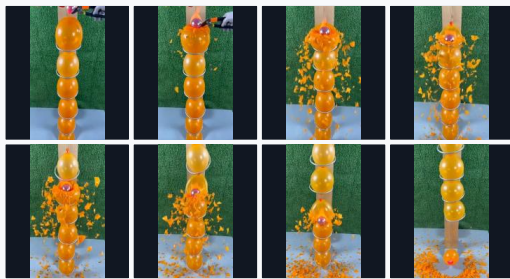
Kling 3.0 Omni

6.0 s



Grok-Image-Video

6.0 s



HappyHorse V1.0

6.0 s



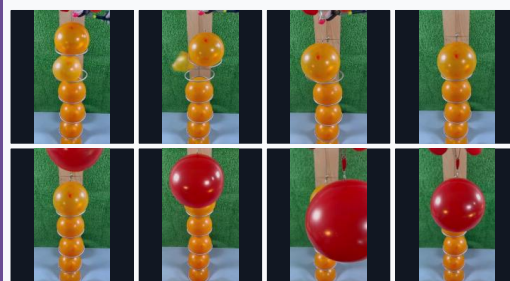
Hailuo 2.3

5.9 s



LTX 2.3 Fast

6.1 s



RAVEN-Eval supplementary material · KFT · Reasoning · ASMR · frames sampled at fixed relative positions · 6/8

Figure 23: A reasoning KFT case collected from an ASMR video. The models are required to predict the rapid sequential bursting of six balloons caused by a falling red-hot iron ball.

KFT Case Example

Reference sequence and six model generations

Non-reasoning

YouTube Source

PROMPT

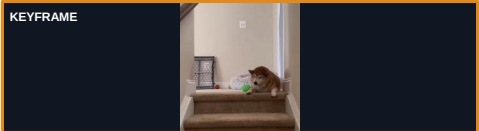
At the start, only a heavier Shiba Inu lies quietly above the stairs. A slimmer Shiba Inu then appears from the upper-left corner of the stairway and finally stops to the left of the first dog.

TASK-SPECIFIC PRIMARY RUBRIC

1. The second, slimmer dog must enter continuously and naturally from the upper-left stair corner and stop to the left of the first dog, without teleportation.

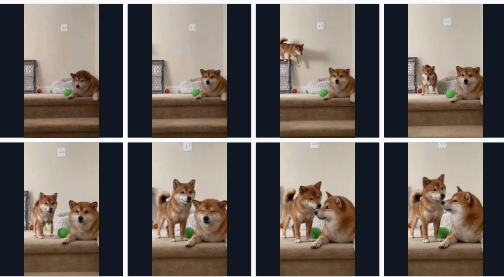
MODEL INPUT · starting keyframe

3.4 s reference



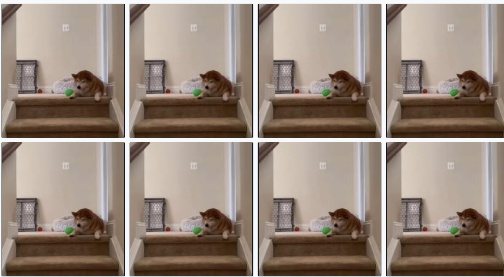
Seedance 1.5 Pro

6.0 s



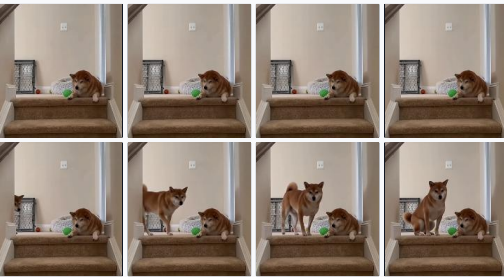
Kling 3.0 Std

6.0 s



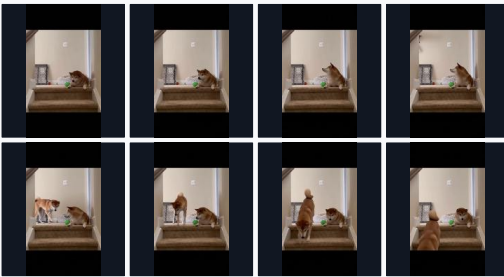
WAN 2.6

6.0 s



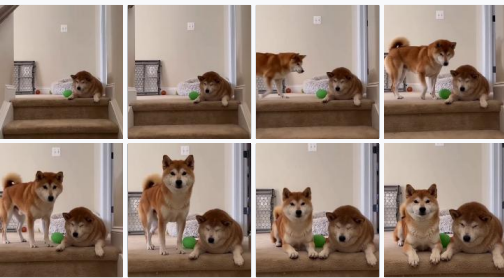
Veo 3.1 Lite

6.0 s



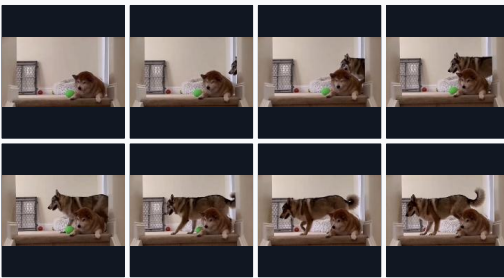
PixVerse C1

6.0 s



LTX 2.0 Pro

6.1 s



RAVEN-Eval supplementary material · KFT · Non-reasoning · YouTube · frames sampled at fixed relative positions · 7/8

Figure 24: A non-reasoning KFT case collected from YouTube. The case evaluates whether the second Shiba Inu enters naturally and stops at the specified position without temporal discontinuities.

KFT Case Example

Reference sequence and six model generations

Non-reasoning

ASMR Source

PROMPT

Show multiple marbles rolling along the tracks. New marbles should continuously emerge from the end of each track, so the number of marbles on every track increases, while each track's marble color remains unchanged.

TASK-SPECIFIC PRIMARY RUBRIC

1. Marbles must roll along all tracks as new marbles continuously emerge and the count on each track increases.
2. The marble color assigned to each track must remain consistent from beginning to end.

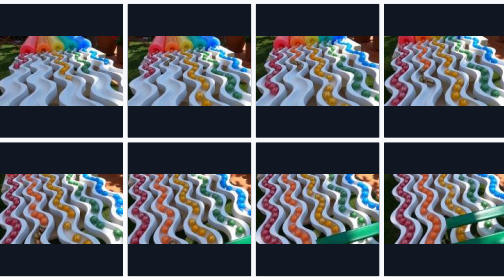
MODEL INPUT · starting keyframe

1.4 s reference



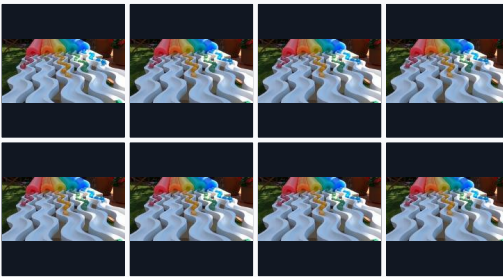
Seedance 2

6.0 s



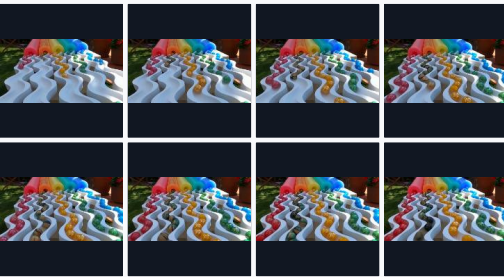
Kling 3.0 Pro

6.0 s



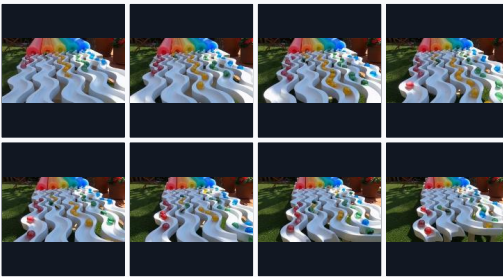
WAN 2.7

6.0 s



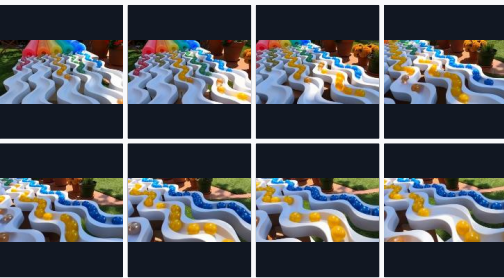
PixVerse V5.5

5.0 s



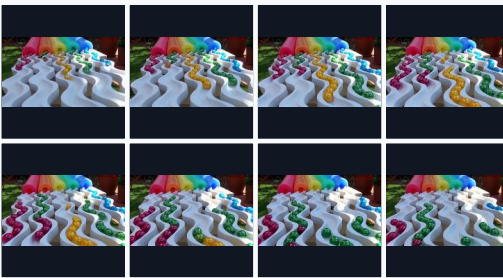
Hailuo 2.3

5.9 s



LTX 2.0 Fast

6.1 s



RAVEN-Eval supplementary material · KFT · Non-reasoning · ASMR · frames sampled at fixed relative positions · 8/8

Figure 25: A non-reasoning KFT case collected from an ASMR video. The case evaluates continuous marble motion, increasing marble counts, and color consistency across the generated sequence.

T2V | T2V Case 1 | SRCC=-0.0872, n=20

Prompt: A busy street scene where at least 10 pedestrians, 3 bicycles, 4 cars, and 5 storefront facades appear at the same time. Pedestrians, bicycles, and cars are all moving. Use a fixed wide-angle street-view camera, emphasize complex multi-subject dynamics, include a slightly swaying roadside sign and a small bag beside a seat, and make motion directions in the foreground, midground, and background...

Reference

N/A

Primary rubric: At least 10 pedestrians, 3 bicycles, 4 cars, and 5 storefront facades are visible; pedestrians, bicycles, and cars move continuously; motion directions at all depths remain readable; subjects keep plausible spatial layout without abnormal overlap or disappearance.

Failure mode: The task is dominated by exact object counts, small props, and multi-depth motion. Some globally strong models missed these local constraints, while lower global models gained rank by simply making the scene crowded and readable.

Best
#1 Veo 3.1 Fast
score 4.19



Middle
#10 Kling 3.0 Pro
score 1.25



Worst
#20 LTX 2.3 Fast
score -5.10



T2V | T2V Case 2 | SRCC=-0.0767, n=20

Prompt: An orchestral rehearsal hall with at least 8 performers, multiple instruments, 4 music-stand areas, and 1 spectator seating area. Performers make slight but synchronized body and arm movements following the conductor. Use a stable wide-angle shot. Include a pencil beside a music stand and a slightly fluttering sheet of music in the front row, keep motion paths continuous, and avoid abrupt jumps or...

Reference

N/A

Primary rubric: At least 8 performers and multiple instruments are visible; performers show slight synchronized motion under conductor cues; the music stands, pencil, and fluttering sheet music remain clear.

Failure mode: The evaluation rewards subtle synchronization and tiny prop visibility. Ranking deviates because broad visual quality is less important here than stable wide-shot structure, identity retention, and small-object fidelity.

Best
#1 Wan 2.6
score 3.07



Middle
#10 LTX 2.3 Pro
score 0.04



Worst
#20 PixVerse V5.5
score -3.58



T2V | T2V Case 3 | SRCC=-0.0647, n=20

Prompt: A port loading area where at least 10 workers, 2 forklifts, 3 handcars, and 5 groups of container edges appear at the same time. Workers, forklifts, and handcars are all moving, and at least one natural crossing or occlusion occurs among moving subjects. Spatial relations should remain readable, with no random drift or identity confusion.

Reference

N/A

Primary rubric: The port scene contains workers, forklifts, handcars, and containers; at least 10 workers and multiple moving vehicles are visible; moving subjects naturally cross or occlude one another.

Failure mode: The task emphasizes dense logistics geometry and occlusion handling. Task scores favor models that preserve container layout and vehicle-worker relations, which differs from the global leaderboard's average preference.

Best
#1 Veo 3.1 Fast
score 9.30



Middle
#10 LTX 2.0 Pro
score -0.18



Worst
#20 Wan 2.6
score -7.61



Figure 26: T2V failure cases.

KFT | KFT Case 1 | SRCC=-0.0072, n=18

Prompt: Show the short-term movement of many vehicles in a complex traffic hub. The vehicle group should move slowly and continuously. **Reference** relative positions and shapes of all roads, as well as the approximate appearance of every vehicle, must remain consistent.

Primary rubric: The traffic hub should contain a large group of slowly moving vehicles; road positions and shapes and the approximate vehicle appearances must remain consistent across the video.

Failure mode: This task is strongly tied to preserving reference geometry. Models that over-animate or alter the road layout fall in task rank, even if they are strong on the overall KFT leaderboard.



Best
#1 Wan 2.6
score 5.70



Middle
#9 veo3.1-fast
score 0.97



Worst
#18 PixVerse V6
score -8.37



KFT | KFT Case 2 | SRCC=0.1022, n=18

Prompt: Based on the principle of communicating vessels, infer the final equilibrium water level in the large bucket and the small bottle, and generate the full water-level change process. The liquid color and the shapes of the bucket and bottle must remain consistent.

Primary rubric: The water level in the small bottle should slowly rise until it visually matches the level in the large bucket; the liquid color and container shapes must remain consistent.

Failure mode: The ranking is driven by a local physical reasoning event rather than general generation quality. Models that maintain shape but fail the water-level equilibrium are penalized heavily.



Best
#1 LTX 2.0 Pro
score 4.88



Middle
#9 PixVerse C1
score 0.62



Worst
#18 Kling 3.0 Omni
score -4.22



KFT | KFT Case 3 | SRCC=0.2120, n=20

Prompt: From the maze shown in the reference image, accurately generate the motion of nine balls rolling through a sloped maze under gravity. Collisions between balls and maze obstacles should look physically realistic. Balls starting from different positions should show distinct motion states, while their count, colors, and shapes stay unchanged.

Primary rubric: Nine balls should move through the sloped maze according to gravity; collisions with obstacles should be realistic; different starting positions should lead to distinct trajectories; ball count, color, and shape remain constant.

Failure mode: The task depends on object counting and multi-body collision dynamics. It can reward models with better local physics and identity preservation even when their global KFT rank is low.



Best
#1 Kling 3.0 Omni
score 4.00



Middle
#10 Kling 3.0 Pro
score 1.46



Worst
#20 LTX 2.0 Fast
score -10.52



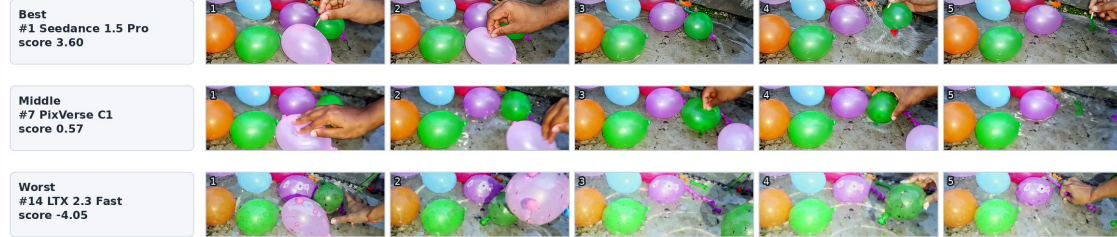
Figure 27: KFT failure cases.

FLT | FLT Case 1 | SRCC=-0.1780, n=14

Prompt: Show the process of first puncturing the lower-right pink balloon and then puncturing the rightmost green balloon. After each puncture, show water ripples and their effect on nearby balloons, making the surrounding balloons shake.

Primary rubric: The lower-right pink balloon should be punctured before the rightmost green balloon; water ripples after popping should affect nearby balloons and make them shake.

Failure mode: The task is sensitive to exact event order and target colors. Global leaders can drop if they produce a plausible pop but choose the wrong balloon sequence or miss secondary water-wave effects.

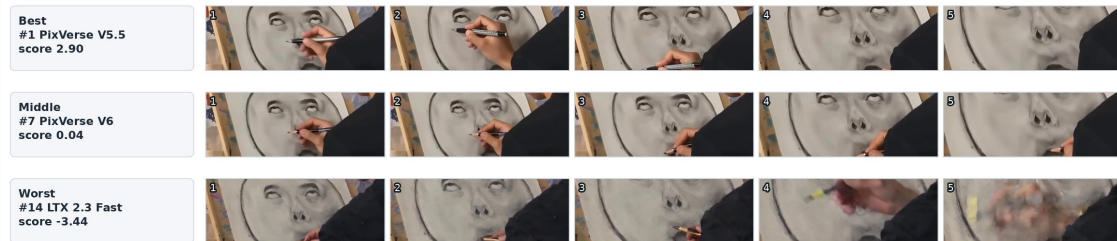
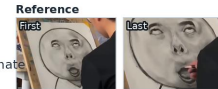


FLT | FLT Case 2 | SRCC=-0.1560, n=14

Prompt: The male student holds a pen in his right hand and continuously refines the drawing. The hand should show relatively high-frequency position changes.

Primary rubric: The male student's right hand should keep drawing and refining the artwork with frequent visible hand-position changes.

Failure mode: The task focuses on subtle localized hand motion. Rankings deviate because models that preserve the scene but under-animate the hand lose task-specific credit.



FLT | FLT Case 3 | SRCC=0.0725, n=14

Prompt: Show a needle rubbing back and forth on the middle yellow balloon several times without bursting it. Then the hand pushes the needle tip hard enough to puncture and burst the balloon. Show the yellow balloon moving during the burst and the burst causing nearby balloons to move.

Primary rubric: The needle should first rub the middle yellow balloon without bursting it, then puncture it; the burst should move the yellow balloon and induce motion in surrounding balloons.

Failure mode: The task requires delayed causality: rubbing must not pop the balloon, but the later puncture must. Many models collapse this sequence or miss the induced surrounding motion.



Figure 28: FLT failure cases.

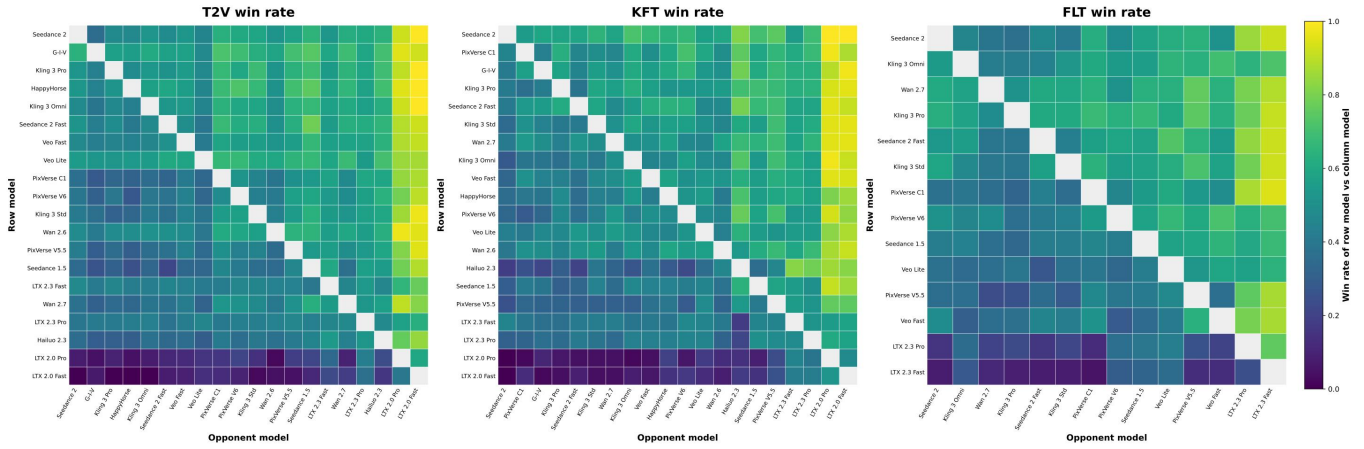


Figure 29: Pairwise win-rate heatmaps for the T2V, KFT, and FLT tasks, computed from all available LMM judge results. Each cell shows the win rate of the row model against the column model, where a merged pairwise sample is counted only when the same judge provides valid forward and reversed labels. Ties and inconsistent forward–reverse judgments are counted as 0.5 wins for both models. Diagonal cells are left blank. For FLT, missing off-diagonal pairs are filled by interpolation for visualization only.

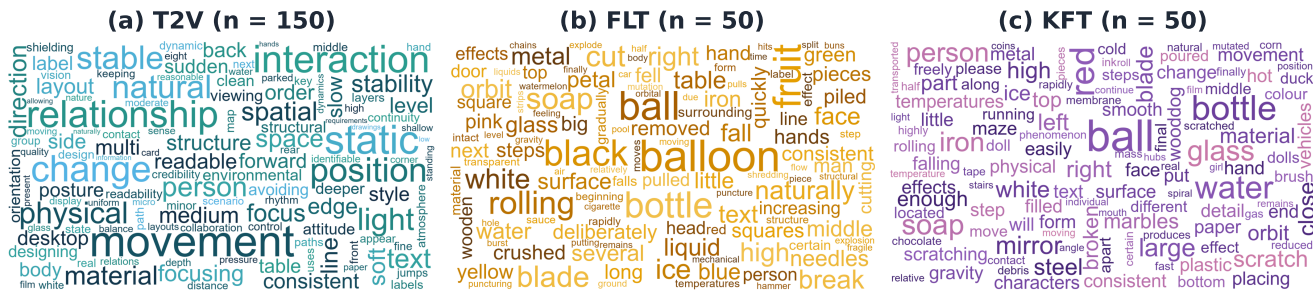


Figure 30: Prompt word clouds for the three evaluation tasks: (a) text-to-video generation (T2V), (b) first–last-frame-to-video generation (FLT), and (c) keyframe-to-video generation (KFT). Word size indicates frequency in the corresponding translated and preprocessed prompt corpus.