FingER: Content Aware Fine-grained Evaluation with Reasoning for AI-Generated Videos

Rui Chen
chenrui.chen@alibaba-inc.com
AMAP, Alibaba Group
Beijing, China

Lei Sun
ally.sl@alibaba-inc.com
AMAP, Alibaba Group
Beijing, China

Jing Tang
guangyu.tj@alibaba-inc.com
AMAP, Alibaba Group
Beijing, China

Geng Li
xiaofeng.lg@alibaba-inc.com
AMAP, Alibaba Group
Beijing, China

Xiangxiang Chu
chuxiangxiang.cxx@alibaba-inc.com
AMAP, Alibaba Group
Beijing, China

Abstract

Recent advances in video generation have posed great challenges in the assessment of AI-generated content, particularly with the emergence of increasingly sophisticated models. The various inconsistencies and defects observed in such videos are inherently complex, making overall scoring notoriously difficult. In this paper, we emphasize the critical importance of integrating fine-grained reasoning into video evaluation. We propose **FingER**, a novel entity-level reasoning evaluation framework that first automatically generates **F**ine-grained **E**ntity-level questions, and then answers those questions by a **R**easoning model with scores, which can be subsequently weighted summed to an overall score for different applications. Specifically, we leverage LLMs to derive entity-level questions across five distinct perspectives, which (i) often focus on some specific entities of the content, thereby making answering or scoring much easier for MLLMs, and (ii) are more interpretable. Then we construct a FingER dataset, consisting of approximately 3.3k videos and corresponding 60k fine-grained QA annotations, each with detailed reasons. Based on that, we further investigate various training protocols to best incentivize the reasoning capability of MLLMs for correct answer prediction. Extensive experiments demonstrate that a reasoning model trained using GRPO with a cold-start strategy achieves the best performance. Notably, our model surpasses existing methods by a relative margin of 11.8% on GenAI-Bench and 5.5% on MonetBench with only 3.3k training videos, which is at most one-tenth of the training samples utilized by other methods. Our codes and datasets have been released.

CCS Concepts

• **Computing methodologies → Scene understanding**.

Keywords

Fine-grained VQA, Reasoning Model, T2V Generation, MLLMs, RL

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755102

ACM Reference Format:
Rui Chen, Lei Sun, Jing Tang, Geng Li, and Xiangxiang Chu. 2025. FingER: Content Aware Fine-grained Evaluation with Reasoning for AI-Generated Videos. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755102

1 Introduction

Recent advancements in Text-to-Video (T2V) generative models [3, 5, 49] have demonstrated significant progress in producing visually appealing and content-rich videos. For instance, post-Sora models such as Kling have shown the ability to generate high-resolution videos that closely adhere to textual prompts. However, these models often produce localized artifacts, inconsistencies, and violations of physical laws. These issues highlight the necessity for the development of robust and reliable quality assessment methods for AI-generated video content.

Early research on evaluating AI-generated videos has primarily relied on feature-based metrics, such as the Frechet Video Distance (FVD) [31] and optical flow-based methods like RAFT [30]. While these methods effectively assess overall visual quality and dynamic characteristics, they fall short in capturing nuanced aspects that require deeper semantic understanding and fine-grained reasoning. To address these limitations, recent studies have introduced MLLMs for more comprehensive evaluations. For example, VideoScore [12] proposes a framework that evaluates five distinct aspects of video quality using an MLLM to assign scores ranging from 1 to 4. VisionReward [41] aligns video generation with human perception by formulating predefined judgment questions and fine-tuning a video-based MLLM to compute weighted scores. Similarly, LiFT [35] learns a reward model that provides reasons and scores across multiple aspects to align the generation model with human preferences. Despite these advancements, two key challenges persist:

(i) **Inadequacy of Fine-grained Video Reasoning**: Although advanced generative models have significantly improved global visual quality by reducing issues such as blurriness and flickering, they still exhibit localized spatiotemporal inconsistencies, distortions, unnatural artifacts, and violations of physical laws, especially in scenarios involving complex motion or multiple entities. For instance, Fig. 1 (a) shows a video generated by PixVerse that, despite its high overall visual appeal, contains a noticeably deformed hand in a localized area. This example underscores the need for more

Figure 1: Advanced generation models often exhibit localized defects while maintaining overall visually appealing, as illustrated in (a), which requires fine-grained in-depth understanding. (b) and (c) show that even with detailed instructional prompts and entity-level questions, GPT-4o still fails to identify this hand deformation. (d) shows the effectiveness of our work by integrating reasoning model with fine-grained entity-level questions.

fine-grained and context-aware reasoning capabilities in video understanding, moving beyond superficial visual pattern recognition to incorporate temporally grounded and semantically rich analysis.

(ii) **Domain Gap in AI-Generated Videos**: Current state-of-the-art MLLMs struggle to capture the intrinsic characteristics of AI-generated videos, even with well-defined prompts. As illustrated in Fig. 1 (b) and (c), GPT-4o misidentifies the deformed hand in a video and assigns a high score based on misleading explanations. This issue is primarily attributed to the domain gap between the training data used by MLLMs and the unique features of AI-generated videos. In essence, AI-generated videos can deceive MLLMs in certain latent feature spaces. Bridging this gap requires a high-quality dataset of AI-generated videos. Moreover, developing strategies to enhance the generalization of MLLMs to AI-generated videos remains an open challenge.

Inspired by the Question Generation and Answering (QG/A) framework [7] and recent reasoning works [8, 24, 25, 50] that demonstrate a significant self-emergence of complex cognitive reasoning abilities induced by Deepseek-R1 [11], we argue that incorporating fine-grained reasoning abilities would significantly enhance the video quality assessment. In this paper, we propose Fing**ER**, a novel framework that first decomposes the overall evaluation into fine-grained entity-level questions and then answers these questions with corresponding scores by a reasoning model, which is fine-tuned on our high-quality dataset using GRPO with a cold-start initialization. Specifically, we employ five distinct aspects as defined in VideoScore [12], including text-to-video alignment, temporal consistency, factual consistency, dynamic degree, and visual quality. By deriving such fine-grained entity-level questions, our framework not only enables the model to explicitly focus on specific characteristics of certain entities, thereby facilitating a more fine-grained understanding, but also enhances interpretability through these structured QA pairs.

Based on these questions, we prompted several strong MLLMs [15, 29] to provide answers. However, we observed that these models struggle to provide correct answers, particularly in aspects like factual consistency. As stated before, we attribute this to the lack of high-quality AI-generated video datasets and the inadequate reasoning capabilities of current models. Therefore, we curated a fine-grained AI-generated video reasoning dataset, **FingER-Instruct-60k**, which consists of 3.3k AI-generated videos sourced from advanced generation models like Kling, Luma, Vidu [3], PixVerse, CogVideoX [42], etc. For each video, we generate fine-grained questions and annotate them with "Yes/No". To ease human labor and also reduce potential errors, we leverage MLLMs to generate detailed reasoning explanations given each question and its answer. (Note that, while MLLMs often struggle to answer these questions correctly, they demonstrate higher possibilities of producing coherent reasoning when the answer is explicitly provided, suggesting the presence of underlying reasoning capabilities.) These generated reasons were subsequently re-checked and refined by human annotators to ensure accuracy and quality. At last, we collected 60k fine-grained QA annotations with high-quality detailed reasons.

To enhance the video reasoning capabilities, we choose Qwen2.5-VL [1], and explore multiple training protocols on our dataset, including directly training with answers, training with reasons, zero GRPO training, and GRPO training with a cold-start initialization. Our experiments reveal that integrating high-quality reasons can largely increase the performance along with the interpretability, and GRPO with cold-start can further enhance its performance, especially in dimensions that require in-depth understanding. We also test our reasoning model in a zero-shot manner on public benchmarks, and still consistently achieve state-of-the-art performance.

In summary, we propose an entity-level quality assessment framework with strong reasoning and generalization capabilities. To the best of our knowledge, our work is the first to introduce entity-level reasoning into the quality assessment of AI-generated videos.

Our contributions can be summarized as follows:

- **Novel Evaluation Approach.** We propose a novel evaluation approach Fing**ER**, designed for practical AI-generated video quality assessment. It comprises an entity-level question generation module and a video reasoning model that provides corresponding scores. By emphasizing fine-grained reasoning, our approach effectively addresses localized defects in AI-generated videos that require in-depth understanding and significantly enhances interpretability.
- **Fine-grained Reasoning Dataset.** We present a new dataset for AI-generated video reasoning, containing 3.3k videos and 60k entity-level QA annotations sourced from advanced generation models. Each QA pair is annotated with detailed

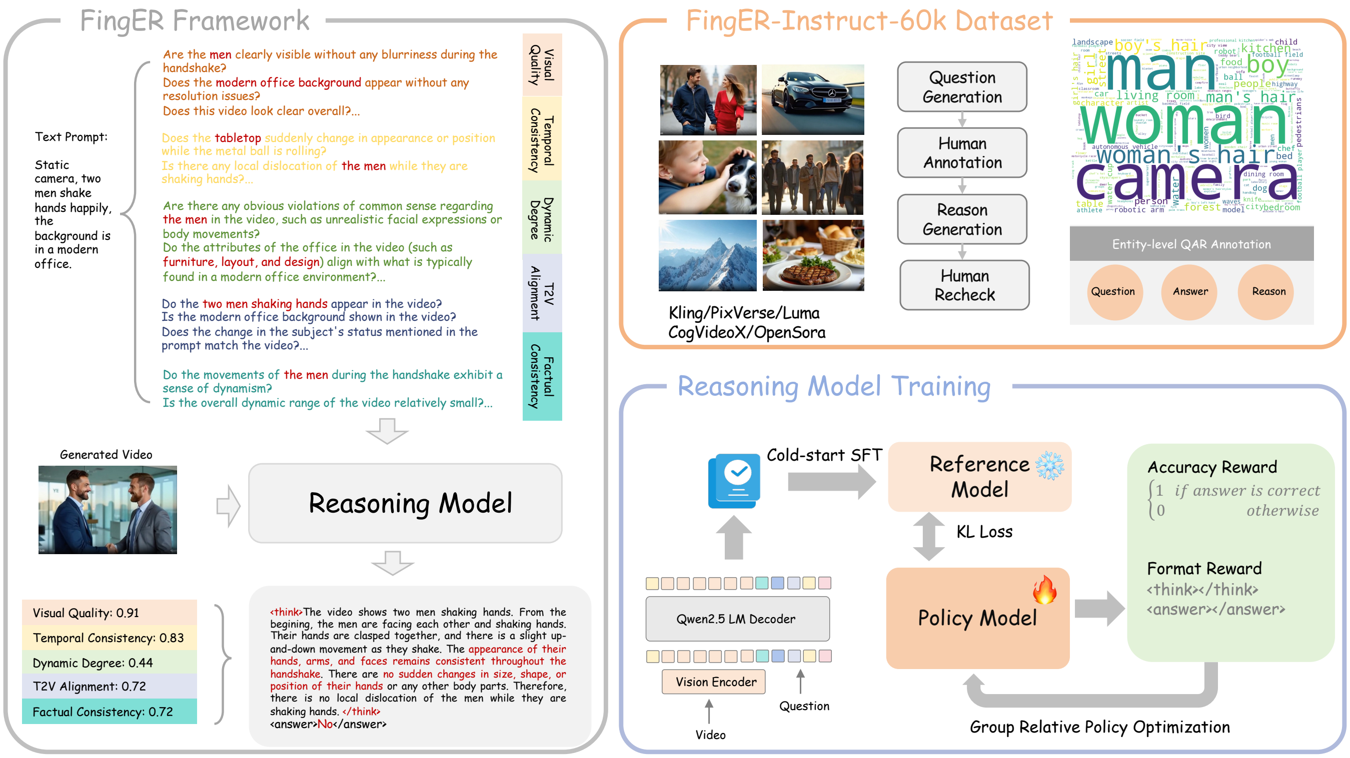

Figure 2: The overview of our proposed FingER framework, including (a) the evaluation pipeline, (b) FingER-Instruct-60k dataset curation, and (c) GRPO training of our reasoning model.

reasons. This dataset aims to further advance research in this field.

- **Enhanced Training Protocols.** We explore multiple training protocols to enhance the fine-grained video reasoning capabilities of MLLMs. Notably, we are the first to introduce GRPO training into AI-generated video quality assessment, which proves to be highly effective in improving both reasoning and generalization abilities
- **Strong Performance**. Extensive experiments demonstrate the effectiveness of our approach. We achieve state-of-the-art performance on public benchmarks using only one-tenth of the training videos, thereby highlighting the superior generalization capability of our model.

2 Related Work

2.1 Video Quality Assessment

Early approaches relied on feature-based metrics, such as Fréchet Video Distance (FVD) [31], Inception Score (IS) [26], and CLIP-Sim [38]. And benchmark works like EvalCrafter [23] and VBench [14, 47] introduced comprehensive evaluation frameworks with 18 and 16 metrics, respectively. However, these methods fall short in assessing deep semantic understanding or aligning with human perception.

With the rapid advancement of MLLMs [1, 6, 9, 29, 46], increasing studies have explored to leverage their capabilities to facilitate image/video quality evaluation [7, 13, 20, 21, 39]. Inspired by DSG [7], which uses question generation/answering (QG/A) for interpretable assessment, T2VScore [40] adopted a QA framework for T2V alignment. T2VQA [16] introduced the T2VQA-DB dataset, comprising 10k videos annotated with Mean Opinion Scores (MOS), and trained a transformer-based model to predict these scores. Similarly, VideoScore [12] proposed a larger dataset across five dimensions and employed a MLLM for scoring. VMBench [22] introduced perception-aligned motion metrics to evaluate motion quality. While these methods predict scores or labels, they often overlook the reasoning behind assessments, limiting their effectiveness. Our work distinguishes itself by incorporating entity-level reasoning for evaluating advanced generation models reliably.

Another line of research focuses on reward models for improving generative models via Reinforcement Learning from Human Feedback (RLHF), such as Diffusion-DPO [33], VisionReward [41] and UnifiedReward [34, 36]. While these efforts target generative model optimization, our work emphasizes practical video quality evaluation, we expect it is able to further benefit the generation models using RLHF in future work.

2.2 Reasoning Inference in Large Models

Reasoning inference aims to emulate human-like thinking processes by forming the final answer through a Large Language Model (LLM). Specifically, to answer a given question, an LLM is required to think divergently and record the thinking processes, which are subsequently referenced when formulating the final answer. This approach has inspired a variety of research, including prompting-based Chain-of-Thought (CoT) [37], planning-based Graph-of-Thought [4] and Tree-of-Thought [43] processing, reward methods [18], and supervised fine-tuning (SFT) datasets with sufficient context [44]. Notably, DeepSeek-R1 [11] integrates specific prompts with reinforcement learning (RL), enabling the model to first generate the thinking process before producing the final answer. This method allows for supervised fine-tuning with a small amount of annotated data containing thinking processes, followed by reinforcement learning fine-tuning on more data without thinking processes. A very recent approach [2, 8] proposes a highly simplified reinforcement learning framework and demonstrates its validity across several benchmarks.

3 Method

3.1 Entity-level VQA Framework

For Text-to-Video (T2V) generation task, user input prompt is the only key instruction for generative models to understand and generate content that well-aligned with user's intent. To perform entity-level quality assessment of AI-generated videos, we start from understanding the user's prompt through extracting entities, attributes, and actions within itself. Inspired by DSG [7] in Text-to-Image (T2I) evaluation, we also utilize closed-source Large-Language-Model (LLM) to perform textual understanding and the following entity extraction. As shown in Fig. 2, we provide abundant in-context learning (ICL) [37] examples from different video generation scenarios and formulate the final input for GPT-4o [15], in which way we can harvest more steady entity extraction results.

With entities extracted from the user's prompt, we generate entity-level questions from **five** distinct video quality assessment dimensions, including *visual quality*, *text-to-video alignment*, *temporal consistency*, *factual consistency*, and *dynamic degree*. For each dimension, we provide a detailed explanation followed by several key points, formulating the context information when prompting the LLM. We also prepare adequate **entity-level** in-context learning examples, which are summarized from videos with and without obvious artifacts or hallucinations. In this way, we can help the LLMs to better understand which question should be asked when coping with a specific entity along with the given assessment dimension. In short, we break down the granularity of fine-grained video quality assessment from multi-dimensional level to entity-level. And the intuition behind entity-level question generation is that we hope fine-grained question/answering can guide the MLLM to focus on understanding the correlation between entity-level textual description and its corresponding visual appearance based on the video content.

After the entity-level question generation procedure, our fine-tuned MLLM answers the above questions with a simple *"Yes"* or *"No"*, along with a detailed reasoning process explaining why the answer is that. Learning the logical reasoning process is critical for model performance improvements, as detailed in the experiment Sec. 4.4. The outputted reason can also be useful when conducting practical video quality assessment, which is more interpretable and user-friendly. To formulate a final score representing the overall quality of AI-generated videos, we start by calculating the probability of the answer token ("Yes" or "No") for each entity-level question to represent the **entity-level score**. Since there are multiple "Yes" and "No" with different formats but similar meanings in the vocabulary of our MLLM, we first gather the token set for "Yes" and "No". In this paper, we take $["Yes", "yes", "YES", ""Yes", " Yes"]$ as the token set for answer "Yes", and $["No", "no", "NO", ""No", " No"]$ for answer "No", denoted by T_Y and T_N, respectively. With logits from the answer token, we extract all the logit whose token id is within the token set, and apply softmax over $T_Y \cup T_N$, as illustrated in Eq. 1. Then, given the entity-level question q, we can get the answer's probability $P(No\,|\,q)$ and $P(Yes\,|\,q)$ with a simple sum up.

$$\begin{aligned} P(No \mid q) &= \sum_{i=1}^{n} Softmax(x_i)\,,\ x_i \in T_N\,; \\ P(Yes \mid q) &= \sum_{j=1}^{m} Softmax(y_j)\,,\ y_j \in T_Y\,. \end{aligned} \tag{1}$$

Instead of directly using the derived probability as the entity-level score, we still need the judgment on whether the question is positive or negative. For example, given the question *"Do the attributes of the table in the video (such as size, shape, and material) align with real-world characteristics?"* from the factual consistency dimension, it is apparent that the factual consistency of the assessed video goes up with a positive *"Yes"* answer. We define this type of question as a positive one, and vice versa. We denote the status of an entity-level question with q_{stat}, if q_{stat} equals 1, it means that the question is positive; otherwise, the question is negative.

$$S_{entity} = \begin{cases} P\,(No \mid q), & if\ q_{stat}\ =\ 0\,; \\ P\,(Yes \mid q), & if\ q_{stat}\ =\ 1\,. \end{cases} \tag{2}$$

With the aforementioned preparations setup, we propose our entity-level score S_{entity}, which correlates positively with the quality of the assessed video. When the entity-level question is positive, we use the probability of the "Yes" answer $P(Yes\,|\,q)$ to represent the score it can gain. And we utilize the probability of the "No" answer $P(No\,|\,q)$ if the question is negative, as illustrated in Eq. 2. In short, our intuition behind this design is that as long as the video quality goes up with which answer, we calculate our entity-level score based on that answer's probability. Then, we utilize entity-level question/answering pairs that are under the same quality assessment dimension to formulate our **dimension-level score** S_{dim}. To be specific, we simply calculate the linear summation of multiple answers' probability S_{entity}, as illustrated in Eq. 3.

$$S_{dim} = \frac{1}{N} \sum_{i=1}^{N} S_{entity} i \tag{3}$$

In the end, we derive the **overall-level score** $S_{overall}$ with the weighted average of five distinct dimension scores S_{dim} in Eq. 4.

$$S_{overall} = \sum_{i=1}^{5} w_i \,.\, S_{dim} i \tag{4}$$

3.2 Entity-level Dataset with Reasoning

In this section, we introduce the construction pipeline of our entity-level instruction tuning dataset, named Fing**ER**-**Instruct-60k**.

3.2.1 ***Prompt and T2V Model Selection.*** Based on VideoGen-Eval [45] dataset, our instruction tuning dataset is composed of 420 diverse text prompts and 3.3k AI-generated videos produced by 8 modern T2V models, including closed-source models: Kling, Luma, PixVerse, Vidu [3], Qingying, and open-sourced models: Mochi-1 [28], CogVideoX [42], Open-Sora [49]. We utilize all 420 text prompts from the T2V session [45], which cover a diverse range of complex scenarios, including human-centric activities, material and spatial relationships, as well as animal and text generations. These prompts are derived from real-life user inputs. As for the T2V model selection, we denote models that understand and obey most of the common sense and physical laws, and generate time-consistent videos without obvious temporal distortions as the high-quality model. We select the generative models uniformly based solely on the quality of their generated videos, spanning from high-quality models to average-quality models, for a more diverse training data distribution.

3.2.2 ***Entity-level Question Generation and Annotation.*** Our multi-dimensional entity-level question generation starts with understanding users' input prompts and extracting the entities within. We use GPT-4o [15] for prompt understanding and entity extraction, with abundant in-context learning examples provided. Then, we perform the entity-level question generation for our five distinct assessment dimensions. For each entity, we prompt the LLM with task introduction, assessment dimension explanation with several key points to focus on, user's input prompt, the extracted entity, and the most important in-context learning examples. And we extract the generated questions with regular expression matching.

For data annotation, we engaged 10 professional annotators to complete the task of annotating 60k question/answer pairs. Inter-annotator agreement was ensured through multiple rounds of small-scale pilot annotations, and the entire process took approximately one month to complete.

3.2.3 ***Reasoning Generation and Verification.*** We employ the powerful closed-source MLLM [29] to generate the initial version of the reasoning process. Specifically, we prompt the MLLM with the assessment dimension explanation, user prompt, in-context learning examples, and the entity-level question along with its human-annotated result. An interesting finding is that when the MLLM is provided with the correct answer to the entity-level question, the generated reasoning process for explaining the answer is more reasonable than when directly generating the answer and its reason. Rather than using the MLLM-generated reasoning process directly, we conduct thorough human verification to ensure the quality of our reasoning training data.

3.3 Instruction Tuning and GRPO Training

We use Qwen2.5-VL-7B-Instruct [1] as our base model and apply supervised fine-tuning, SFT with reasoning and reinforcement learning on it.

3.3.1 ***Supervised Fine-Tuning.*** We directly train the base model on Fing**ER-Instruct-60k**, the response of model only contains "Yes" or "No" answer following the next token prediction paradigm. It means the model only needs to learn to predict the correct answer without any reasoning process. The loss function is Cross-Entropy Loss:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{5}$$

3.3.2 ***Supervised Fine-Tuning with Reasoning.*** We also train base model on Fing**ER-Instruct-60k**, but the difference compared to Supervised Fine-Tuning is the model needs to learn predicting the correct answer within $< answer > ... < /answer >$ tag and its reasoning processes within $< think > ... < /think >$ tag. We apply prompt engineering on the input tokens to reach this difference. The loss also contains the gap of reasoning processes and the gap of answers.

3.3.3 ***GRPO Training.*** We employ GRPO [27] to enhance reasoning inference performance, exploring two protocols: (i) Zero-GRPO, which relies solely on reinforcement learning without initial supervised data; and (ii) GRPO with cold-start Supervised Fine-Tuning, which combines initial supervised learning with subsequent reinforcement optimization.

Zero-GRPO. Zero-GRPO is an exploratory attempt that is initiated directly from Qwen-2.5-VL [1] and uses RL to implicitly improve reasoning abilities without annotated reasons. For each video-question pair, we first sample a group of outputs $\{o_1, o_2, ..., o_G\}$ by old policy $\pi_{\theta_{old}}(o_i|v, q)$, v denotes the video that needs to be evaluated, q denotes the question for each entity and dimension. Then update the policy model π_θ by minimizing the following loss.

$$\begin{aligned} \mathcal{L}_{GRPO}(\theta) = -\mathrm{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta old}(O|v, q)] \\ \frac{1}{G}\sum_{i=1}^{G}\Bigg(min\Bigg(\frac{\pi_\theta(o_i|v, q)}{\pi_{\theta old}(o_i|v, q)} * Adv_i, \\ clip\Bigg(\frac{\pi_\theta(o_i|v, q)}{\pi_{\theta old}(o_i|v, q)}, 1-\epsilon, 1+\epsilon\Bigg) * Adv_i\Bigg) \\ + \beta\mathcal{D}_{KL}(\pi_\theta||\pi_{ref})\Bigg) \end{aligned} \tag{6}$$

$$\mathcal{D}_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|v, q)}{\pi_\theta(o_i|v, q)} - log\frac{\pi_{ref}(o_i|v, q)}{\pi_\theta(o_i|v, q)} - 1 \tag{7}$$

β denotes the coefficient of Kullback-Leibler Divergence [17] between base model and policy model, ϵ denotes the threshold of clip. Adv_i is the advantage which is the normalization of a group of rewards $\{r_1, r_2, ..., r_G\}$ computed from outputs within each group:

$$Adv_i = \frac{r_i - Mean\{r_1, r_2, ..., r_g\}}{Std\{r_1, r_2, ..., r_G\}} \tag{8}$$

r_i is composed of two reward functions:

$$r_i = r_{accuracy_i} + r_{format_i} \tag{9}$$

$$r_{accuracy_i} = \begin{cases} 1.0 & \text{if } answer_i = GT_i \\ 0.0 & \text{else} \end{cases} \tag{10}$$

$$r_{format_i} = \begin{cases} 1.0 & \text{if } o_i \text{ includes correct format} \\ 0.0 & \text{else} \end{cases} \tag{11}$$

Correct format means the output o_i contains two tags:
$< answer > ... < /answer >$ and $< think > ... < /think >$.

"Yes" or "No" token only appears within the answer tag, and the reasoning process only appears within the reason tag.

GRPO with cold-start Supervised Fine-Tuning. DeepSeek-R1 demonstrated that fine-tuning on an annotated dataset with reasoning processes before applying reinforcement learning (RL) yields better performance than directly using RL [11]. We adopt this approach in our supervised fine-tuning model. The sole difference between Zero-GRPO and GRPO with cold-start Supervised Fine-Tuning lies in the base model: the latter is initialized from a model pre-trained on annotated data containing reasoning processes.

4 Experiments

4.1 Datasets and Evaluation Metrics

4.1.1 ***Datasets.*** We split 185 generated videos (around 5% of the whole data) with 3.5k entity-level questions from 5 distinct quality assessment dimensions to formulate our **FingER-test** dataset. Regarding the public benchmarks, we adopt the popular GenAI-Bench [19] and recently released MonetBench [41] for performance evaluation. GenAI-Bench contains 800 unique text prompts paired with 4 T2V models, and each generated video has MOS (Mean Opinion Scores) annotated by 3 annotators. MonetBench consists of 1000 different text prompts, each paired with 2 T2V models. Each pair of videos is generated with the same prompt but different video generation models. MonetBench annotates the video pair with human preferences, including "win", "lose", and "tie" options.

4.1.2 ***Evaluation Metrics.*** We report the accuracy (Acc) of "Yes" or "No" answers, the Pearson linear correlation coefficient (PLCC), and the Spearman rank correlation coefficient (SRCC) on our proposed FingER-test dataset. We evaluate our models with and without token probability calculation, denoted by (*w/ prob*) and (*w/o prob*) in Tab. 1 and Tab. 2. Following previous works in [12, 21], we utilize the SRCC and the PLCC for evaluating model's performance on GenAI-Bench. And we use pairwise accuracy as the metric for human preference evaluation on MonetBench and report *tau* and *diff*, followed [10, 41].

4.2 Implementation Details

Based on Qwen-2.5-VL-7B [1], we fine-tune our model with the following experiment settings: learning rate of 5.0e-6, global batch size of 32, video input fps (frame-per-second) is set to 2, and video maximum input resolution is set to 448 × 448 pixels. We utilize LLaMA-Factory [48] as our supervised fine-tuning (SFT) codebase. We perform SFT on our proposed FingER-Instruct-60k dataset for 2 epochs with 8 NVIDIA H20 GPUs, and the training steps are the same for the model trained with extra reasoning process. As for the settings of our reinforcement learning (RL) experiments, we employ Huggingface-TRL [32] as our RL fine-tuning tool with following hyper-parameters to implement GRPO: $\beta = 0.04$, and the number of group $G = 16$, $\epsilon = 0.2$, $\mu = 1$, the initial learning rate of RL is 5.0e-7. We train Zero-GRPO and GRPO with cold-start for 2k steps on 4 NVIDIA H20 GPUs. And w_i is set to 0.2 for $S_{overall}$ in Eq. 4.

4.3 Zero-shot Performance on FingER-test

We report the zero-shot performance of Qwen2.5-VL across five dimensions on our dataset. Through ablations on resolution, frame rate (fps), and evaluation granularity, we reveal the capabilities of the base model to handle different dimensions, and further demonstrate the crucial importance of integrating entity-level evaluation.

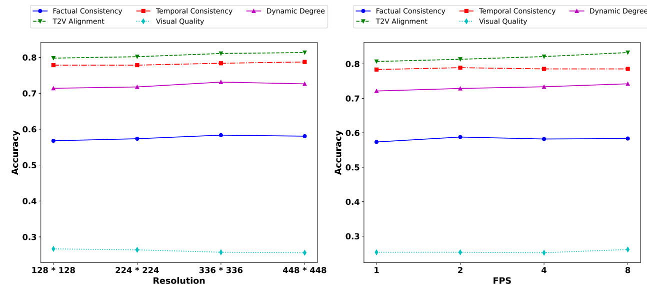

Figure 3: Zero-shot performance on five distinct assessment dimensions with different input resolution and fps.

Increasing resolution and fps leads to slight improvements. Fig. 3 illustrates the accuracy across five dimensions when prompted with entity-level questions. We can see that the accuracy curves show slight improvements with increasing resolutions or frame rates (fps), albeit at a significant computational cost. These results suggest that resolution and fps are not the primary factors of performance enhancement. Consequently, for efficiency we adopt 448 × 448 pixels and 2 fps as the default settings for subsequent zero-shot and supervised fine-tuning (SFT) experiments.

Performance varies significantly across different dimensions. As shown in Fig. 3, the zero-shot accuracy for visual quality is exceptionally low at 26.1%, while factual consistency achieves 57.6%. In contrast, dimensions like text alignment show higher accuracy at 80.59%, likely due to the base model's inherent capabilities from pre-training on caption data. We believe that the notably low accuracy in visual quality is primarily attributed to misalignment from AI-generated videos, and the main challenges still lie in dimensions requiring in-depth reasoning, such as factual consistency, temporal consistency, and text alignment, which will be further demonstrated in the following section.

Integrating entity-level evaluations brings a substantial performance gain. To validate the efficacy of our entity-level QA framework, we conduct experiments across three evaluation granularities: **overall level**, **dimension level**, and our proposed **entity level**, as detailed in Tab. 1. The **overall level** (*1st row*) prompts the model with an overall assessment rating from 1 to 4, accompanied by detailed evaluation criteria, while the **dimension level** (*2nd row*) prompts model to rate each dimension from 1 to 4, which are then averaged to get a final score. The results of our proposed **entity-level** (*3rd and 4th rows*) are reported with and without a probability calculation strategy introduced in Sec. 3.1, and furthermore, we instruct the model to provide explanatory reasoning along with answers (*last two rows*). Compared to the entity-level framework, both the overall and dimension levels exhibit substantial performance degradation across all dimensions, indicating that fine-grained evaluation substantially enhances the model's performance. It is worth noting that incorporating explanatory reasoning does not bring improvements, revealing the inherent limitations of the base model in understanding AI-generated videos.

Table 1: Correlation between model Zero-shot answer and human reference on FingER-test

Method	Visual Quality	Temporal	Dynamic Degree	Text Alignment	Factual	Overall
Qwen2.5-VL	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC
Overall Level	—	—	—	—	—	-/30.68/29.27
Dimension Level	-/35.06/35.54	-/16.05/17.06	-/14.81/14.09	-/33.68/32.62	-/13.86/12.28	-/52.32/61.14
Entity (w/o prob)	25.33/1.85/5.22	78.72/83.26/83.91	72.87/51.04/48.98	81.6/70.68/73.44	58.34/51.03/53.27	66.50/80.86/83.71
Entity (w/ prob)	25.33/40.60/40.94	**78.72**/84.51/**85.44**	72.87/**56.48**/**56.85**	**81.6**/**74.09**/**76.49**	**58.34**/**57.45**/**58.67**	**66.50**/**81.23**/**85.26**
+Reason (w/o prob)	45.71/**49.97**/49.61	77.65/83.12/83.89	75.21/54.30/52.87	81.08/73.24/75.31	40.51/17.43/23.55	63.96/73.40/79.15
+**Reason** (w/ prob)	**45.71**/46.29/**49.64**	77.65/**84.60**/83.89	**75.21**/48.88/52.80	81.08/72.38/75.35	40.51/29.27/23.50	63.96/73.29/79.18

Table 2: Correlation between SFT/RL model answer and human reference on FingER-test (Z-GRPO means Zero-GRPO)

Method	Visual Quality	Temporal	Dynamic Degree	Text Alignment	Factual	Overall
	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC	Acc/SRCC/PLCC
GPT-4o [15]	62.19/56.24/57.93	77.83/78.64/79.13	68.31/54.14/57.02	83.41/72.20/74.33	58.77/48.93/49.51	69.92/81.25/82.36
VideoScore [12]	-/22.80/18.55	-/23.84/26.06	-/9.49/7.18	-/19.18/13.87	-/22.93/18.31	-/20.39/17.68
Qwen2.5-VL [1]	25.33/40.60/40.94	78.72/84.51/85.44	72.87/56.48/56.85	81.6/74.09/76.49	58.34/57.45/58.67	66.50/81.23/85.26
Z-GRPO (w/o prob)	76.01/73.39/70.46	78.01/83.13/83.82	77.93/69.74/68.47	84.46/73.80/75.99	55.21/47.47/50.33	74.51/83.46/86.56
Z-GRPO (w/ prob)	76.01/71.83/71.97	78.01/81.81/83.86	77.93/67.49/68.51	84.46/74.38/76.28	55.21/42.21/50.15	74.51/83.24/86.82
FingER (w/o prob)	83.78/83.48/82.53	83.33/83.13/83.70	83.23/71.37/67.95	82.77/70.94/73.75	72.89/64.12/64.61	81.25/88.87/89.67
FingER (w/ prob)	83.78/**85.31**/**85.22**	83.33/86.24/86.99	**83.23**/**77.07**/74.73	82.77/73.85/77.98	72.89/70.99/69.26	81.25/90.23/91.41
+Reason (w/o prob)	84.05/81.51/81.00	84.04/85.88/86.63	82.49/69.22/68.22	86.79/77.87/79.77	74.03/67.47/68.41	82.33/89.79/91.64
+**Reason** (w/ prob)	**84.05**/83.85/83.87	**84.04**/86.51/87.09	82.49/76.11/**76.70**	**86.79**/**79.34**/**83.16**	74.03/71.70/70.27	**82.33**/90.31/92.04
+GRPO (w/o prob)	82.30/80.62/78.09	82.98/85.08/85.57	81.63/65.54/64.92	85.88/75.74/77.91	74.04/68.65/70.73	81.41/89.26/91.25
+**GRPO** (w/ prob)	82.30/83.76/83.51	82.98/**86.64**/**87.43**	81.63/75.05/74.68	85.88/78.32/82.63	**74.04**/**71.87**/**72.03**	81.41/**90.43**/**92.41**

Table 3: Zero-shot Evaluation Results on Public Benchmarks

Method	GenAI-Bench[19]		MonetBench[41]	
	SRCC	PLCC	tau	diff
GPT-4o[15]	35.79	36.61	45.70	48.30
Qwen2.5-VL[1]	46.62	44.29	46.70	44.27
VideoScore[12]	42.22	40.62	49.10	54.90
VQAScore[21]	52.70	50.60	56.10	59.50
Zero-GRPO	49.58	44.39	51.30	51.34
FingER	54.13	52.60	53.90	57.31
+ Reason	56.68	**57.25**	57.80	62.07
+ GRPO	**57.03**	56.59	**58.00**	**62.80**

4.4 SFT and RL Performance on FingER-test

In this section, we report the performance of our reasoning model on Fing**ER**-test using different training protocols, including SFT with answers, SFT with reasons, zero GRPO, and GRPO with a cold start. We also provide results using the closed-source model GPT-4o [15] and VideoScore [12] for comparisons, as detailed in Tab. 2. Note that all these results, except for VideoScore [12], are obtained by entity-level evaluations for fair comparisons.

Our model, trained with only answers, demonstrates significant performance improvements over the base model, achieving overall gains of **14.75/9.00/6.15** in Acc/SRCC/PLCC, respectively. Substantial improvements are observed in the dimensions of visual quality, dynamic degree, and factual consistency. Note that the improvement in the text alignment dimension is limited, mainly due to its inherent capabilities derived from pre-training data.

Incorporating additional reasoning during training further boosts the performance, particularly in the dimensions of text alignment, factual consistency, and temporal consistency. For the text alignment dimension, the SFT with reasoning harvests performance gains with **4.02/5.49/5.18** in Acc/SRCC/PLCC. These improvements underscore the importance of in-depth video understanding to achieve higher performance in these dimensions.

We further investigate the reasoning training using RL, which includes two kinds of training procedures: (1) Zero-GRPO, and (2) GRPO initialized with a cold-start from reasoning SFT training. The results presented in Tab. 2 reveal that Zero-GRPO fails to predict correct answers. Upon closer examination of the training process, we identified that the issue stems from the reasoning component. Zero-GRPO generates reasons that resemble captions rather than logical reasoning. In contrast, when GRPO is applied with a cold-start initialization from our reasoning SFT model, it is able to surpass the SFT model with only 1k additional training steps. Among these dimensions, we observed steady performance improvements in the *temporal* and *factual consistency* dimensions, with boosts of **1.15/0.88/2.77** in *factual consistency*. We believe that the reasoning cold-start teaches the model to reason in a rough manner, while GRPO guides it towards adopting reasons with correct answers, thereby incentivizing its reasoning capability.

Moreover, we evaluate the performance on our proposed FingER-test dataset with closed-source MLLM [15] *(1st row)*, and VideoScore

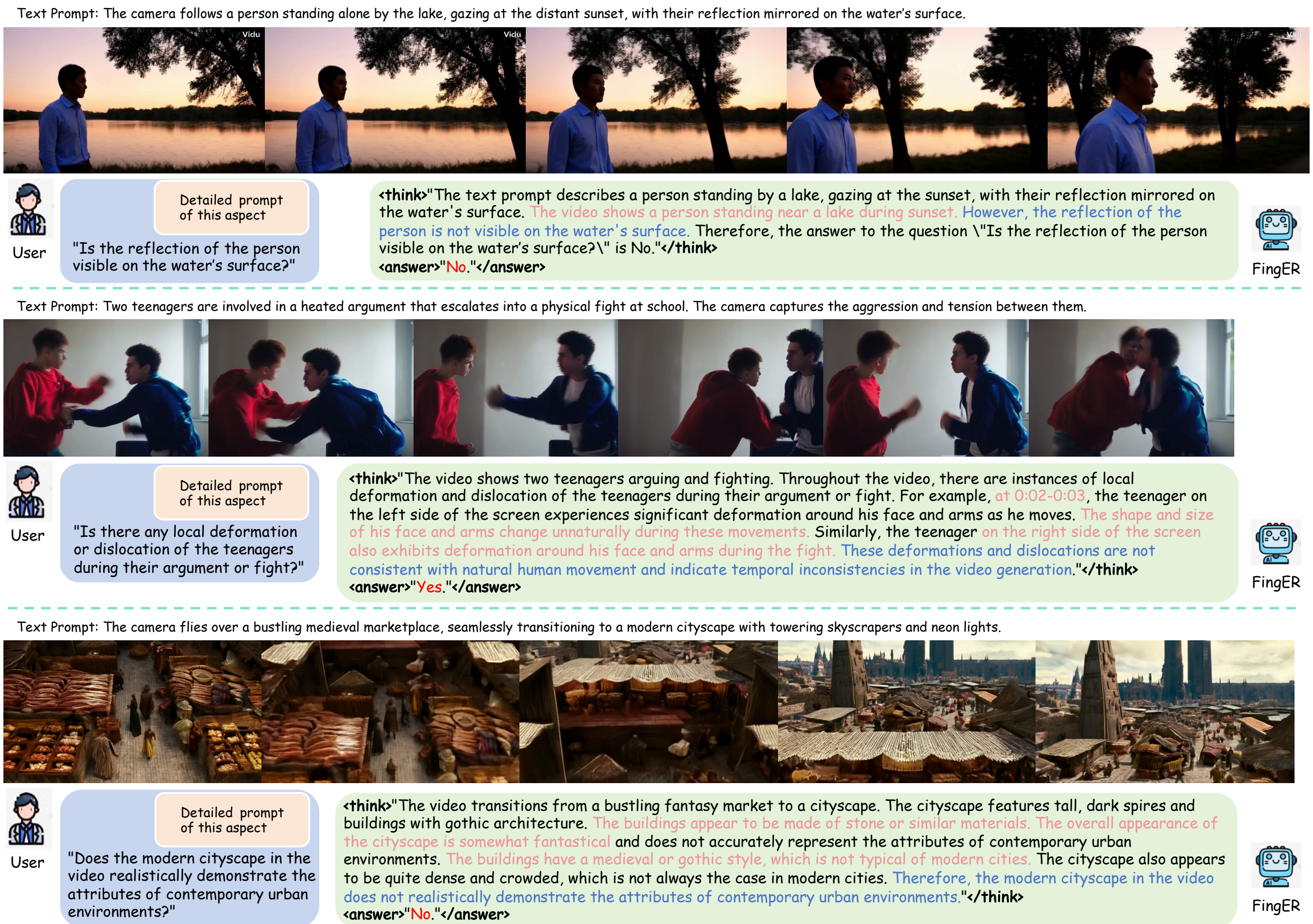

Figure 4: Qualitative results. We show several reasoning results outputted by our GRPO model.

[12] *(2nd row)*, our proposed FingER outperforms those methods with a large margin across all five assessment dimensions.

4.5 Comparison on Public Benchmarks

Tab. 3 demonstrates the consistent improvements achieved by our method on two public benchmarks. We compare our methods with GPT-4o, Qwen2.5-VL and two other approaches. Specifically, with only Yes/No answer prediction, we already outperform all methods on GenAI-Bench, indicating the effectiveness of our fine-grained evaluation framework. Training with reasons and GRPO with a cold-start leads to further improvements with a final 8.21%/11.83% SRCC/PLCC relative performance boost. On MonetBench, without any weight fitting, we just average scores of five dimensions, our method is able to achieve 3.39%/5.55% relative improvements of tau/diff. It is worth noting that VideoScore [12] is trained using 37.6k training videos, while VQAScore [21] utilizes 665k samples, we outperform these methods with only 3.3k training videos without additional training samples from other sources, which is at most one-tenth of the training size adopted by other methods.

5 Conclusion

In this paper, we emphasize the critical importance of integrating fine-grained reasoning into AI-generated video quality assessment, and we propose **FingER**, an entity-level fine-grained quality assessment framework with five distinct evaluation dimensions for AI-generated videos. To bridge the gap between non-AI videos and AI-generated videos, we construct a high-quality dataset, **FingER-Instruct-60k**, which consists of 3.3k videos generated by modern T2V models and 60k entity-level question / answering / reasoning pairs. Based on this dataset, we explore multiple training protocols to best incentivize the model's reasoning capability, including reason SFT, zero GRPO and GRPO with a reasoning cold-start. Extensive experiments demonstrate that by utilizing GRPO training with a cold-start, our method not only achieves the best performance on our dataset, but also outperforms other methods and closed-source models on two public benchmarks. And it is worth noting that we achieve SOTA performance with only 3.3k training samples. Our codes and datasets are available at https://github.com/AMAP-ML/FingER.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. 2025. Univg-r1: Reasoning guided universal visual grounding with reinforcement learning. *arXiv preprint arXiv:2505.14231* (2025).

[3] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233* (2024).

[4] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 17682–17690.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog* 1, 8 (2024), 1.

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.

[7] Jaemin Cho, Yushi Hu, Jason M Baldridge, Roopal Garg, Peter Anderson, Ranjay Krishna, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *ICLR*.

[8] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. 2025. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546* (2025).

[9] Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766* (2024).

[10] Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12914–12929.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[12] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. 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.

[13] Hailang Huang, Yong Wang, Zixuan Huang, Huaqiu Li, Tongwen Huang, Xiangxiang Chu, and Richong Zhang. 2024. MMGenBench: Fully Automatically Evaluating LMMs from the Text-to-Image Generation Perspective. *arXiv preprint arXiv:2411.14062* (2024).

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

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[16] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu. 2024. Subjective-aligned dataset and metric for text-to-video quality assessment. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7793–7802.

[17] Solomon Kullback. 1951. Kullback-leibler divergence. *Tech. Rep.* (1951).

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629* (2024).

[19] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Emily Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Genai-bench: A holistic benchmark for compositional text-to-visual generation. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

[20] Mingxing Li, Rui Wang, Lei Sun, Yancheng Bai, and Xiangxiang Chu. 2025. Next Token Is Enough: Realistic Image Quality and Aesthetic Scoring with Multimodal Large Language Model. *arXiv preprint arXiv:2503.06141* (2025).

[21] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*. Springer, 366–384.

[22] Xinran Ling, Chen Zhu, Meiqi Wu, Hangyu Li, Xiaokun Feng, Cundian Yang, Aiming Hao, Jiashu Zhu, Jiahong Wu, and Xiangxiang Chu. 2025. VMBench: A Benchmark for Perception-Aligned Video Motion Generation. *arXiv preprint arXiv:2503.10076* (2025).

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

[24] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783* (2025).

[25] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785* (2025).

[26] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

[27] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).

[28] Genmo Team. 2024. Mochi 1. https://github.com/genmoai/models.

[29] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[30] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.

[31] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. *ICLR 2019 Workshop DeepGenStruct* (2019).

[32] Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. https://github.com/huggingface/trl.

[33] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8228–8238.

[34] Yibin Wang, Zhimin Li, Yuhang Zang, Chunyu Wang, Qinglin Lu, Cheng Jin, and Jiaqi Wang. 2025. Unified multimodal chain-of-thought reward model through reinforcement fine-tuning. *arXiv preprint arXiv:2505.03318* (2025).

[35] Yibin Wang, Zhiyu Tan, Junyan Wang, Xiaomeng Yang, Cheng Jin, and Hao Li. 2024. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814* (2024).

[36] Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. 2025. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236* (2025).

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[38] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. 2021. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806* (2021).

[39] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. 2024. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *International Conference on Machine Learning*. PMLR, 54015–54029.

[40] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. 2024. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781* (2024).

[41] Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. 2024. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059* (2024).

[42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

[43] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[44] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387* (2025).

[45] Ailing Zeng, Yuhang Yang, Weidong Chen, and Wei Liu. 2024. The dawn of video generation: Preliminary explorations with sora-like models. *arXiv preprint arXiv:2410.05227* (2024).

[46] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 46, 8 (2024), 5625–5644.

[47] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. 2025. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755* (2025).

[48] Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 400–410.

[49] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404* (2024).

[50] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. 2025. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv preprint arXiv:2503.05132* (2025).