VCapsBench: A Large-scale Fine-grained Benchmark for Video Caption Quality Evaluation

Shi-Xue Zhang[1 3*], **Hongfa Wang**[2 3*], **Duojun Huang**[3*], **Xin Li**[3*], **Xiaobin Zhu**[1†], **Xu-Cheng Yin**[1]

[1]University of Science and Technology Beijing, China
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[3]Tencent Technology (Shenzhen) Co. Ltd, China
zhangshixue111@163.com

Abstract

Video captions play a crucial role in text-to-video generation tasks, as their quality directly influences the semantic coherence and visual fidelity of the generated videos. Although large vision-language models (VLMs) have demonstrated significant potential in caption generation, existing benchmarks inadequately address fine-grained evaluation, particularly in capturing spatial-temporal details critical for video generation. To address this gap, we introduce the Fine-grained Video Caption Evaluation Benchmark (**VCapsBench**), the first large-scale fine-grained benchmark comprising 5,677 (5K+) videos and 109,796 (100K+) question-answer pairs. These QA-pairs are systematically annotated across 21 fine-grained dimensions (*e.g.*, camera movement, and shot type) that are empirically proven critical for text-to-video generation. We further introduce three metrics (Accuracy (**AR**), Inconsistency Rate (**IR**), Coverage Rate (**CR**)), and an automated evaluation pipeline leveraging a large language model (LLM) to verify caption quality via contrastive QA-pairs analysis. Our benchmark can advance the development of robust text-to-video models by providing actionable insights for caption optimization.

Code — https://github.com/anoycode22/VCapsBench

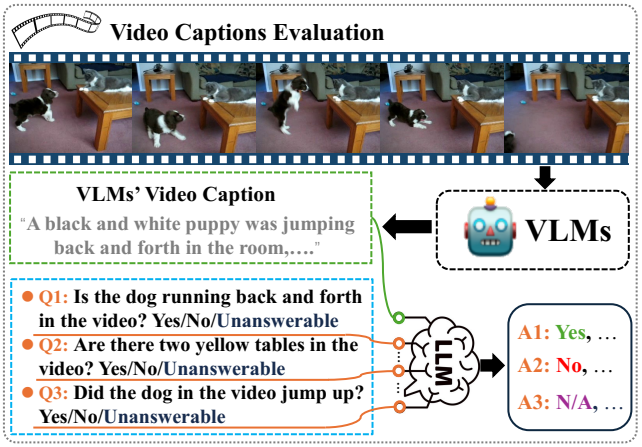

Figure 1: Illustration of video caption evaluation by VCapsBench. Evaluate the detail, comprehensiveness, and accuracy of video captions using "yes-no" question-answer pairs.

Introduction

Recent advances in video understanding (Reid et al. 2024; Liu et al. 2025a; Zhang et al. 2022; Hong et al. 2024; Zhang et al. 2024a; Li et al. 2024a) and generation (Brooks et al. 2024; Bao et al. 2024; Tian et al. 2024; Kong et al. 2024) have been driven by large vision-language models (VLMs). For video comprehension, researchers have extended image-based architectures (*e.g.*, PLLaVA (Xu et al. 2024) and CogVLM2-Video (Hong et al. 2024)) and explored hybrid image-video training approaches (Qwen2-VL (Wang et al. 2024a), LLaVA-OneVision (Li et al. 2025)). Currently, video generation systems such as Sora (Brooks et al. 2024) and HunyuanVideo (Kong et al. 2024) use VLMs for multimodal captioning and prompt engineering. However, current benchmarks (Li et al. 2024b,c; Wang et al. 2025; Fu et al. 2025; Zhou et al. 2024; Wu et al. 2024) struggle to assess the detailed spatio-temporal aspects required for these applications.

In visual generation, existing caption evaluation metrics fall mainly into two categories: reference-based (METEOR (Banerjee and Lavie 2005), BLEU (Papineni et al. 2002), SPICE (Anderson et al. 2016), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)) and reference-free (InfoMetIC (Hu et al. 2023), CLIPScore (Hessel et al. 2021), and TIGEr (Jiang et al. 2019)). Reference-based metrics (Banerjee and Lavie 2005; Papineni et al. 2002; Anderson et al. 2016; Vedantam, Lawrence Zitnick, and Parikh 2015) assess the quality of the captions by comparing them with ground-truth captions. However, these scores are highly dependent on the reference captions' format. Hence, FAIEr (Wang et al. 2021) adopts visual and textual scene graphs for a more robust reference-based evaluation. Reference-free metrics (Hu et al. 2023; Hessel et al. 2021; Jiang et al. 2019) use semantic vectors from the reference image to evaluate the similarity of the caption. These methods falter with concept-dense captions, overwhelmed by numerous concepts. Recent innovations like QACE (Lee et al. 2021), DSG (Cho et al. 2024)], DPG-bench (Hu et al. 2024) and CapsBench (Liu

*These authors contributed equally.
†The corresponding author.

et al. 2024) employ question-answering frameworks to address dense concept evaluation for image captioning. Notably, these methods derived from image description evaluation still have significant measurement blind spots when dealing with the spatio-temporal dynamic elements of video generation scenarios, *e.g.*, camera motion ("slow zoom" vs. "fast pan") or dynamic spatial relationships (object displacement from "left foreground " to "central midground"). In these cases, video generation systems (*e.g.*, Sora) need to ensure accurate text-video alignment.

Current video understanding benchmarks primarily focus on holistic semantic alignment (MVBench (Li et al. 2024b), VideoVista (Li et al. 2024c)) or specific skill evaluation (LVBench (Wang et al. 2025)), neglecting fine-grained spatial-temporal dynamics essential for video generation. This gap becomes critical when evaluating text-to-video systems where caption precision directly impacts visual output. For instance, failing to distinguish between "slow zoom" versus "fast pan" camera motions, or misrepresenting object trajectories from "left foreground" to "central midground". While manual evaluation of such nuances remains impractical at scale, automated metrics also struggle with dynamic semantic alignment due to inherent limitations in traditional evaluation paradigms. To address this problem, it is necessary to re-examine the design principles of video description evaluation metrics.

To address critical gaps in evaluating video caption, we introduce **VCapsBench**, the first large-scale fine-grained benchmark for video caption evaluation. It contains 5,677 diverse videos with 109,796 human-verified questions, allowing fine-grained caption quality evaluation. Specifically, we employ text-based question-answering across 21 categories, including action, camera movements (*e.g.*, zoom, pan, and tilt), object positioning (absolute or relative position), entity and shot type, *etc.*, with "yes", "no", and "unanswerable" ternary judgments. Unlike image-focused caption benchmarks like CapsBench (Liu et al. 2024), our VCapsBench prioritizes temporal continuity through video-specific queries (avg. 19 questions/video) while mitigating LLM hallucination via an "Unanswerable" option. To accurately and comprehensively assess caption quality, we introduce three metrics: Accuracy (AR), Inconsistency Rate (IR), and Coverage Rate (CR). To automate calculating these metrics, we develop an evaluation pipeline that leverages powerful LLMs for objective video caption quality measurement, as illustrated in Fig. 1. In summary, our contributions are as follows:

- We introduce **VCapsBench**, the first large-scale fine-grained benchmark for video caption evaluation, featuring diverse videos (5K+) and QA-pairs (100K+).
- We introduce three metrics: Accuracy (**AR**), Inconsistency Rate (**IR**), and Coverage Rate (**CR**), along with an automated evaluation pipeline to fairly assess the correctness and coverage of video captions.
- We evaluate ten VLMs, including seven advanced open models (Qwen2.5VL, InternVL2.5, and VideoLLaMA3, *etc.*) and three closed model, GPT-4o, Gemini2.5-Pro-Preview, providing a solid reference for the community.

Related Work

Reference-based metric methods (Banerjee and Lavie 2005; Papineni et al. 2002; Anderson et al. 2016; Vedantam, Lawrence Zitnick, and Parikh 2015; Wang et al. 2021) evaluate caption quality by comparing generated captions with ground-truth captions. BLEU (Papineni et al. 2002; Zhou et al. 2023) is a fast, cost-effective, and language-independent metric for machine translation, correlating well with human assessments. METEOR (Banerjee and Lavie 2005) also evaluates machine translation and shows a high correlation with human judgments, significantly outperforming BLEU. CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) measures the similarity of generated sentences against a set of human-written ground-truth sentences, serving as an automatic consensus metric for image description quality. However, these metrics primarily focus on n-gram overlap, which is neither necessary nor sufficient for simulating human judgment. To address these limitations, SPICE introduces a metric based on semantic propositional content over scene graphs, though its scores highly rely on the format of reference captions. Newer methods like FAIEr (Wang et al. 2021) leverage visual and textual scene graphs for a more robust evaluation.

Reference-free metric methods (Hu et al. 2023; Hessel et al. 2021; Jiang et al. 2019; Zhang et al. 2024b) utilize semantic vectors from the reference image to assess caption similarity. InfoMetIC (Hu et al. 2023) is an informative metric for reference-free image caption evaluation to identify incorrect words and unmentioned image regions with fine-grained precision. It provides a text precision score, a vision recall score, and an overall quality score at a coarse-grained level, with the latter showing significantly better correlation with human judgments than existing metrics across multiple benchmarks. CLIPScore (Hessel et al. 2021) employs clip-embedding for robust automatic evaluation of image captioning without references, focusing on image-text compatibility. Although complementary to reference-based metrics emphasizing text-text similarities, CLIPScore is relatively weaker for tasks requiring richer contextual knowledge, e.g., news captions. TIGEr (Jiang et al. 2019) assesses caption quality by evaluating both the representation of image content and the alignment of machine-generated captions with human-generated ones.

Question-based evaluation methods (Lee et al. 2021; Cho et al. 2024; Liu et al. 2024) developed to enhance the assessment of caption quality in visual generation tasks. The QACE framework (Lee et al. 2021) generates questions from captions to evaluate their quality. A similar approach has been proposed for image generation models, where the Davidsonian Scene Graph (DSG) (Cho et al. 2024) organizes questions into dependency graphs, facilitating comprehensive evaluation of text-to-image models. Inspired by DSG and DPG-bench (Hu et al. 2024), Playground v3 (Liu et al. 2024) introduces CapsBench, a benchmark for image captioning that uses "yes-no" question-answer pairs. However, these benchmarks focus solely on image captioning. In contrast, video captioning requires consideration of additional factors such as actions, motion, camera movements, and shot types, which are crucial for text-to-video generation.

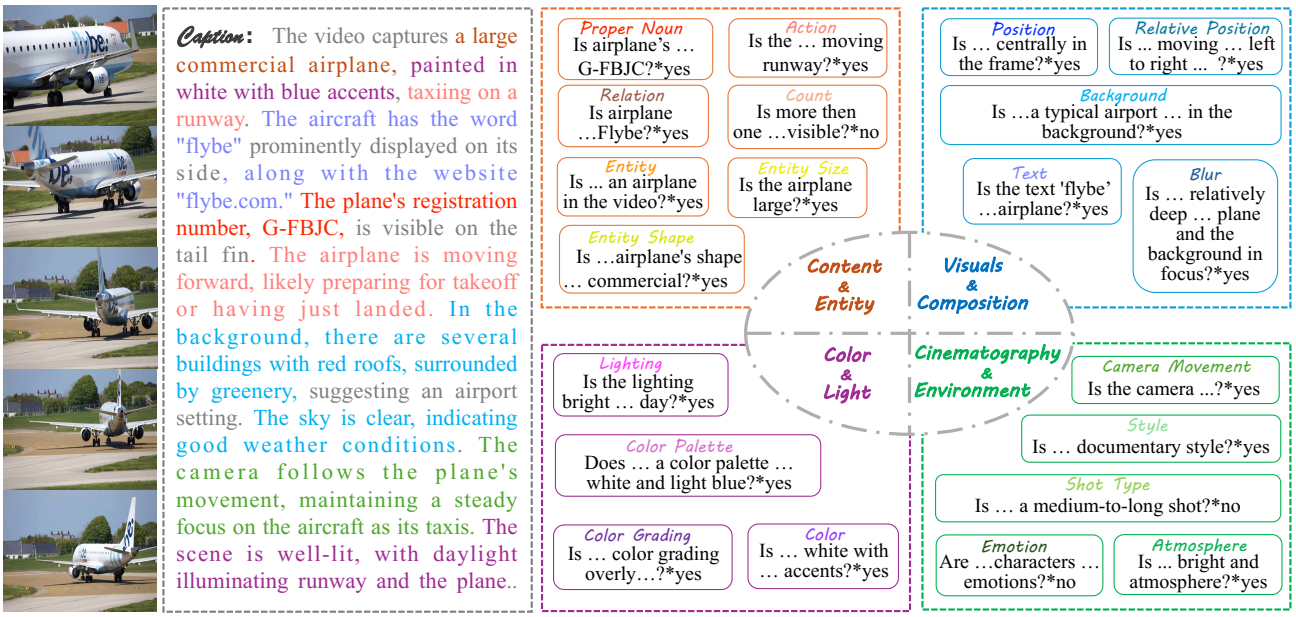

Figure 2: An example of video caption and question-answer pairs in our VCapsBench.

The VCapsBench Benchmark

Dataset Statistics

Data Dimensions. As illustrated in Fig. 2, our video caption evaluation is organized into four primary categories. The first, **Content and Entity**, focuses on core elements through seven subcategories, including action, count, entity, entity size, entity shape, proper noun, and relation. The second, **Visuals and Composition**, pertains to spatial arrangement and visual presentation, comprising five subcategories: position, relative position, text, blur, and background. The third category, **Color and Lighting**, addresses the enhancement of mood and style via four subcategories: color, color palette, color grading, and lighting. Finally, **Cinematography and Atmosphere** integrates filming techniques and artistic expression with five subcategories: camera movement, shot type, style, atmosphere, and emotion. This structured categorization facilitates a statistical analysis of caption quality across different dimensions, thereby providing detailed guidance for optimizing the video understanding capabilities of VLMs.

Data Collection. To fully support the caption evaluation task in video understanding and generation tasks, we prioritized the complexity of aesthetic and content in our data collection for VCapsBench. We sourced videos from 10 publicly available datasets to ensure diversity: Panda-70M (Chen et al. 2024b), Ego4D (Grauman et al. 2022), BDD100K (Yu et al. 2020), Pixabay (Chen et al. 2024a), Pexel (Chen et al. 2024a), VIDGEN-1M (Tan et al. 2024), ChronomicBench (Yuan et al. 2024), FineVideo (Farré et al. 2024), FunQA (Yuan et al. 2024), and LiFT-HRA-20K (Wang et al. 2024b). As illustrated in Fig. 3, we create a diverse collection of 988 high-resolution videos from Panda-70M (Chen et al. 2024b), encompassing a wide range of scenes such as wildlife, cooking, sports, TV shows, gaming, and 3D rendering. These videos often contain complex content and transformations, providing a robust foundation for understanding various real-world scenarios. Additionally, we included 494 high-resolution videos from Pexels and 298 from Pixabay, both renowned for their scenic landscapes and human activities, characterized by high aesthetic quality and detailed imagery. To further ensure data diversity, we sample 1,018 videos from FineVideo (Farré et al. 2024), encompassing 6 major categories and 122 subcategories. Despite their lower resolution (below 640 $\times$360), we balanced this by sampling 333 ultra-high-resolution videos (2K, 3K, and 4K) from VIDGEN-1M (Tan et al. 2024), typically used for training text-to-video models due to their high detail and quality. Our collection was further enriched with videos from Ego4D (Grauman et al. 2022) and BDD100K (Yu et al. 2020) to cover ego-centric human activities and autonomous driving scenarios, ensuring a comprehensive representation of real-world scenes. To assess the VLM model's grasp of object motion and physical laws, we used 1,549 videos from ChronomicBench, spanning categories such as biological, artificial, meteorological, and physical, across 75 subcategories. Additionally, we included 227 videos from FunQA, featuring human-centric content like humorous clips, creative performances, and visual illusions, and 200 synthetic videos from LiFT-HRA-20K.

As shown in Fig. 3, our VCapsBench dataset comprises 5,677 videos from a wide range of scenes, e.g., natural landscapes, animals, human activities, physical phenomena, games, 3D renderings, and synthetic videos (more than 100 subcategories). VCapsBench also features diverse video durations (4 to 16 seconds), resolutions (125 different resolutions), and aspect ratios (87 different ratios). This extensive collection allows for a thorough evaluation of VLM models' understanding and insight across various video types, ensuring a robust assessment of video captioning.

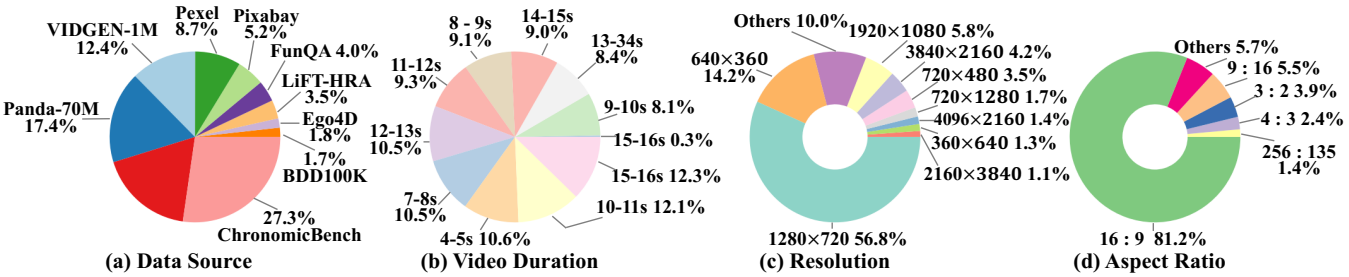

Figure 3: (a) Source distribution; (b) Duration distribution; (c) Resolution distribution (d) Aspect ratio distribution.

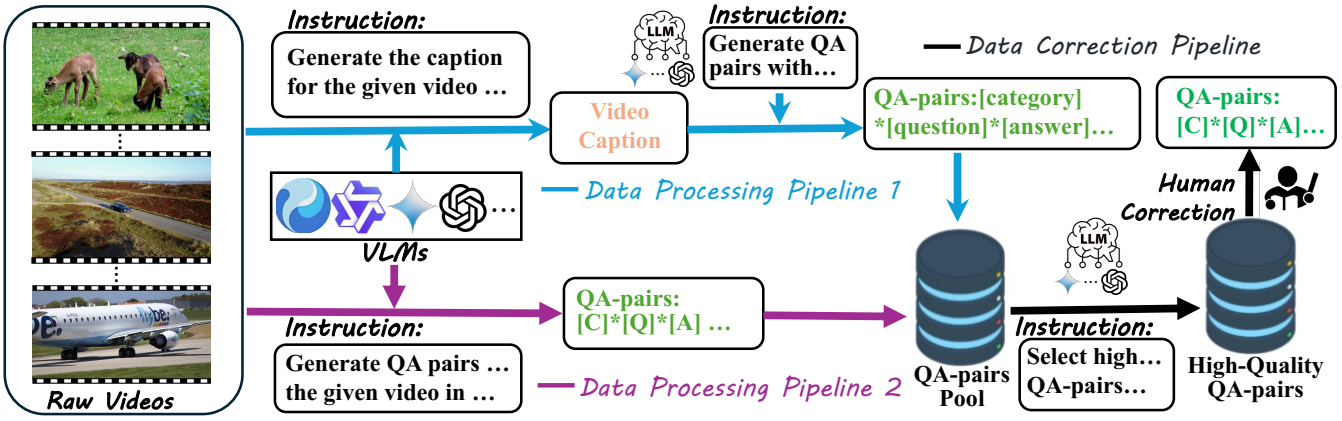

Figure 4: The pipeline of QA-pairs generation, which includes multiple data processing and a data correction pipeline.

Annotation Details

Inspired by DPG-bench (Hu et al. 2024) and CapsBench (Liu et al. 2024), we developed a method to evaluate video caption quality using multi-dimensional question-answer pairs. Based on the raw video, we generate "yes-no" question-answer pairs as annotations. As shown in Fig. 4, our "Data Processing Pipeline" creates these pairs, which are stored in the "QA-pairs Pool". Each pair includes a category, question, and answer, formatted as "[category][question][answer]" for easy processing. To ensure diversity and accuracy, we designed two data production pipelines. The first one uses various VLMs, such as Gemini, Qwen, and HunYuan, to generate detailed video captions from predefined prompts. Large Language Models (LLMs) like GPT-4, Gemini, and Qwen72B then use these captions and additional prompts to create question-answer pairs. The second approach directly utilizes VLMs to generate question-answer pairs from videos and prompts. The outputs from both pipelines are then combined into a comprehensive question-answer pool.

To enhance the quality of the QA-pairs, we established a data correction pipeline. This pipeline takes multiple sets of QA-pairs and captions from the same video as input and uses an advanced LLM (Gemini1.5) with predefined instructions to de-duplicate, filter, and retain high-quality QA-pairs. The instructions guide the LLM to merge similar QA-pairs, filter out those with the same question but different answers, and remove QA-pairs that appear only once (may be invalid).

Human Correction. After generating high-quality candidate QA-pairs, we sample the data according to latitude to ensure that the data focuses more on important latitudes, and then conduct a final quality control process manually. Human reviewers re-examine these QA-pairs, deleting those with unreasonable or incorrect questions and correcting those with erroneous answers. Following this manual revision, we establish a benchmark VCapsBench consisting of 5,677 videos and 109,796 question-answer pairs. As shown in Fig. 5, each video contains between 10 and 27 QA-pairs, with each category comprising between 2,431 and 6,513 QA-pairs. The questions are organized into four major categories and 21 subcategories. Most answers are "yes" providing a clear indication of correctness, while a smaller number of "no" answers help assess whether the video captions contain hallucinations.

Captioning Evaluation

We employ VCapsBench to evaluate video captions produced by various VLMs, like Gemini-1.5. For each test video, an evaluation-capable VLM generates a detailed caption according to specific guidelines and an output schema. This caption, along with each question from the video's QA pairs, is fed into a LLM. As shown in Fig. 1, the LLM responds to each question based on the caption, providing answers in the format "[answer], [reason]." We instruct the LLMs to assess the caption in the following three scenarios:

- **Positive:** The caption accurately describes the relevant content, and the LLM's response aligns with the answer.
- **Negative:** The caption mentions the relevant content, but

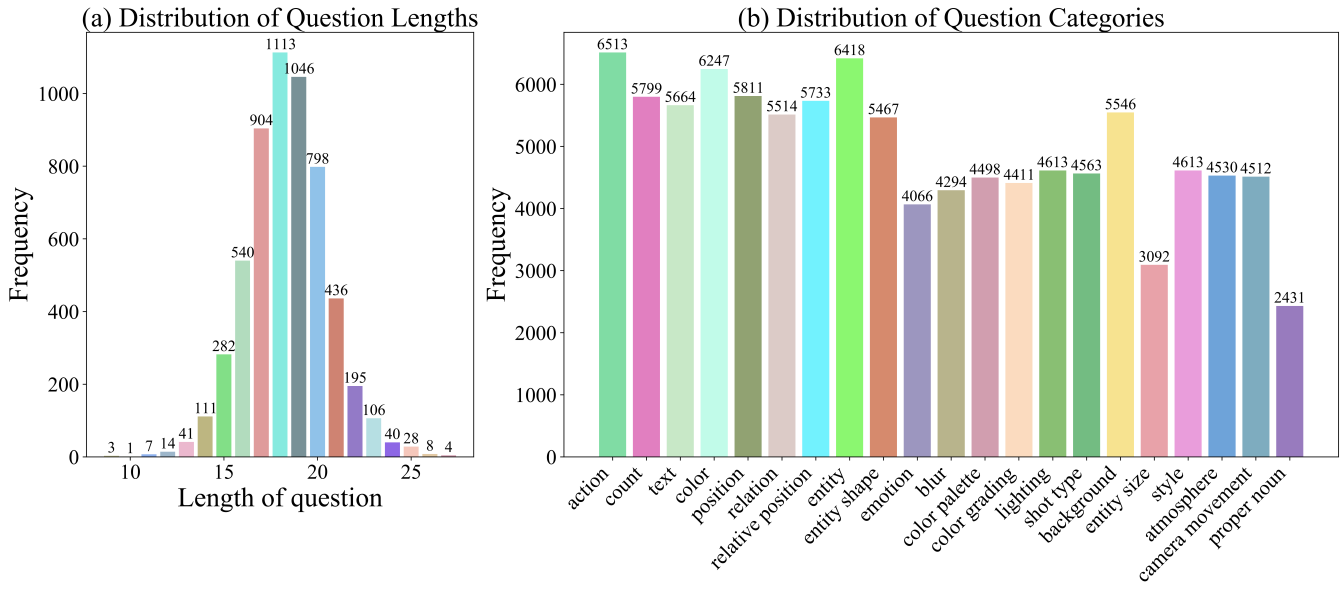

Figure 5: An qnalysis of question word length and categorical distribution.

the LLM's response does not align with the answer.

- **Unanswerable:** The caption does not involve the relevant content for the dimension.

To comprehensively evaluate the quality of the captions, we developed three metrics tailored to these scenarios. The captions' quality is evaluated using the following measure:

$$AR = \frac{N(\text{Positive})}{N(\text{All})}, \tag{1}$$

$$IR = \frac{N(\text{Negative})}{N(\text{Positive}) + N(\text{Negative})}, \tag{2}$$

$$CR = \frac{N(\text{Positive}) + N(\text{Negative})}{N(\text{All})}, \tag{3}$$

Accuracy (AR): This metric evaluates the percentage of "Positive" responses, indicating that the caption correctly describes the relevant content and aligns with the LLM's response. A higher AR signifies a more accurate and reliable caption, reflecting better quality in the LLM's outputs.

Inconsistency Rate (IR): This metric measures the percentage of "Negative" responses among all responses that reference the relevant content (both "Positive" and "Negative"). A lower IR denotes a more accurate caption, indicating that when the caption involves relevant content, it is more likely to be consistent with the real video content.

Coverage Rate (CR): This metric evaluates the total percentage of "Positive" and "Negative" responses, reflecting whether the caption contains the relevant content, independent of the LLM's response consistency. A higher CR indicates greater caption richness, as it means the caption includes more relevant content from the video.

Experiment

Experimental Setup. We evaluated several popular open-source vision-language models (VLMs), including Qwen2VL (Wang et al. 2024a), Qwen2.5VL (Bai et al. 2025), InternVL2.5 (Bai et al. 2025), LLaVA-Video (Zhang et al. 2024c), NVILA (Liu et al. 2025b), and VideoLLaMA3 (Zhang et al. 2025), alongside the closed-source Gemini-2.5 and GPT-4o accessed via API. All models were prompted identically to generate detailed video descriptions across 21 dimensions. For better evaluation, we adopted two powerful LLMs (*e.g.*, Gemini-2.5-Pro-Preview and GPT-4.1) as TextQA experts to answer VCapsBench questions based on the generated descriptions. Following the Playground v3 protocol (Liu et al. 2024), each caption was queried three times, and a consensus response was obtained to minimize output variability and ensure consistency.

Main Results

Table 1 and Fig. 6 presents the accuracy rate (AR), inconsistency rate (IR), and coverage rate (CR) of various VLM methods across multiple dimensions. Overall, Gemini-2.5-Pro-Preview achieves the best or second-best performance in nearly all categories. Notably, it attains the highest AR and CR in most dimensions, such as "Background" and "Camera Movement", reaching 92.72% and 92.09%, respectively, which demonstrates its strong multimodal understanding capabilities. Meanwhile, its IR remains low, indicating good consistency in its responses. GPT-4o also shows competitive results in certain aspects (such as "Entity Shape" and "Lighting"), but its overall performance is still slightly behind the Gemini series. In contrast, smaller models like LLaVA-Video-7B and Qwen2VL-7B lag significantly in all metrics, especially in complex scenarios and fine-grained attributes (*e.g.*, "Blur" and "Camera Movement"). In summary, as model scale and multimodal capabilities increase, VLMs exhibit

	Methods	Content & Entity							Color & Light			
		Proper Noun	Action	Relation	Count	Entity	Entity Size	Entity Shape	Lighting	Color Palette	Color Grading	Color
AR ↑	LLaVA-Video-7B	40.7	57.1	51.7	54.9	34.0	66.2	26.6	57.6	61.5	23.8	56.0
	Qwen2VL-7B	39.2	56.3	52.9	57.2	32.3	64.2	25.6	69.9	55.6	30.6	50.8
	VideoLLaMA3-7B	38.9	58.5	52.6	56.3	33.5	64.3	25.6	60.6	59.0	25.2	54.1
	NVILA-8B	40.0	44.2	48.0	58.0	37.3	63.7	29.3	72.6	66.0	42.4	59.6
	InternVL2.5-8B	38.6	53.9	48.7	59.6	36.8	60.9	29.4	81.3	75.4	60.4	55.8
	Qwen2.5VL-7B	47.0	63.9	57.7	62.4	41.8	69.4	34.9	83.6	73.9	51.3	59.8
	Qwen2.5-VL-72B	48.8	67.3	63.0	67.3	45.4	72.5	37.8	88.4	75.3	61.3	62.7
	GPT-4o	51.1	69.5	66.3	70.1	51.4	77.5	47.9	88.0	82.5	68.2	67.1
	Gemini2.5-Pro-Flash	60.2	76.2	73.5	79.6	64.0	85.3	59.6	**90.1**	86.6	**69.1**	79.4
	Gemini-2.5-Pro-Preview	**61.4**	**78.1**	**75.2**	**80.5**	**64.4**	**86.5**	**62.2**	88.9	**88.1**	67.0	**80.0**
IR ↓	LLaVA-Video-7B	13.1	15.1	11.3	13.1	3.8	8.6	8.5	**5.6**	11.1	**4.1**	13.2
	Qwen2VL-7B	8.4	13.8	10.3	11.7	3.8	8.6	**6.1**	7.5	12.9	5.7	13.4
	VideoLLaMA3-7B	9.9	17.9	12.0	12.6	3.5	8.4	7.6	6.2	11.5	4.4	12.8
	NVILA-8B	12.0	20.	15.0	15.5	4.5	12.4	9.6	8.1	15.7	6.5	17.8
	InternVL2.5-8B	14.6	19.0	16.4	14.2	5.5	14.0	10.2	8.1	15.1	8.6	16.0
	Qwen2.5VL-7B	8.7	14.7	12.4	11.4	**3.2**	9.3	8.4	6.9	11.5	5.5	13.9
	Qwen2.5-VL-72B	7.9	15.0	11.6	11.7	5.0	9.0	8.3	6.9	13.4	5.3	15.2
	GPT-4o	7.5	**11.3**	**9.6**	10.6	3.9	6.5	8.5	6.3	11.7	7.3	13.2
	Gemini2.5-Pro-Flash	6.8	12.6	9.6	9.8	5.2	**5.4**	8.4	6.4	11.3	7.0	**11.2**
	Gemini-2.5-Pro-Preview	**6.6**	13.1	9.8	**9.1**	4.29	5.9	8.6	7.1	**10.2**	5.8	12.2
CR ↑	LLaVA-Video-7B	46.9	67.3	58.3	63.2	35.4	72.4	29.1	61.0	69.1	24.8	64.5
	Qwen2VL-7B	42.8	65.4	58.9	64.7	33.5	70.3	27.3	75.4	63.8	32.5	58.6
	VideoLLaMA3-7B	43.2	71.3	59.7	64.4	34.7	70.2	27.7	64.6	66.7	26.4	62.1
	NVILA-8B	45.4	55.4	56.5	68.5	39.1	72.8	32.4	79.0	78.3	45.3	72.2
	InternVL2.5-8B	45.2	66.5	58.3	69.5	38.9	70.8	32.8	88.4	88.8	66.0	66.5
	Qwen2.5VL-7B	51.4	74.8	65.9	70.4	43.1	76.4	38.1	89.8	83.5	54.3	69.4
	Qwen2.5-VL-72B	52.9	79.2	71.3	76.2	47.8	79.6	41.2	95.0	87.0	64.7	74.0
	GPT-4o	55.3	78.3	73.4	78.3	53.5	82.9	52.3	93.9	93.4	73.6	77.2
	Gemini2.5-Pro-Flash	64.6	87.2	81.3	88.3	**67.5**	90.1	65.0	**96.2**	97.6	**74.2**	89.4
	Gemini-2.5-Pro-Preview	**65.7**	**89.9**	**83.3**	**88.6**	67.3	**91.9**	**68.1**	95.7	**98.0**	71.0	**91.0**

	Methods	Visuals & Composition					Cinematography & Environment					
		Position	Relative Position	Background	Text	Blur	Style	Camera Movement	Shot Type	Emotion	Atmosphere	**ALL**
AR ↑	LLaVA-Video-7B	34.3	33.0	69.3	43.8	13.5	66.6	23.7	38.1	48.5	82.0	47.7
	Qwen2VL-7B	35.3	32.4	68.7	34.8	15.2	71.5	16.3	32.7	55.2	83.6	47.5
	VideoLLaMA3-7B	35.4	32.5	65.4	38.1	17.0	71.7	39.3	39.5	53.3	80.4	48.5
	NVILA-8B	40.7	33.4	69.7	38.0	26.4	74.1	17.0	39.4	53.5	81.0	49.7
	InternVL2.5-8B	39.1	32.6	70.9	37.3	36.8	86.6	51.3	50.5	58.8	86.6	54.9
	Qwen2.5VL-7B	45.0	38.6	77.2	45.0	30.7	88.6	51.9	50.9	63.5	89.5	58.7
	Qwen2.5-VL-72B	49.8	43.0	79.0	45.6	36.3	92.0	66.5	59.4	67.9	89.9	63.1
	GPT-4o	57.2	49.0	83.1	46.8	48.0	92.1	55.6	63.0	62.5	**90.8**	66.5
	Gemini2.5-Pro-Flash	70.4	63.6	87.6	**63.7**	67.3	**96.1**	70.8	74.4	73.0	86.2	75.6
	Gemini-2.5-Pro-Preview	**70.6**	**66.1**	**88.0**	63.5	**68.8**	95.8	**74.9**	**75.7**	**77.1**	88.3	**76.7**
IR ↓	LLaVA-Video-7B	**10.5**	11.8	7.5	17.9	**10.1**	**1.6**	24.1	**12.2**	**3.7**	**3.7**	10.1
	Qwen2VL-7B	11.2	**10.4**	8.8	13.5	14.5	2.7	19.2	13.2	5.2	6.3	**9.7**
	VideoLLaMA3-7B	12.3	12.2	7.9	14.9	12.2	2.2	23.4	14.7	3.7	4.5	10.5
	NVILA-8B	16.2	16.5	10.3	17.9	18.6	4.5	23.6	22.5	6.3	6.1	13.1
	InternVL2.5-8B	15.3	14.7	8.7	19.1	21.9	4.9	26.2	20.0	6.1	5.6	13.4
	Qwen2.5VL-7B	12.8	13.5	8.4	12.3	19.8	2.7	22.8	16.4	5.5	5.1	10.6
	Qwen2.5-VL-72B	13.3	13.5	8.0	12.6	20.7	3.4	21.0	17.0	5.7	5.1	11.0
	GPT-4o	12.4	12.4	6.2	12.9	22.3	3.6	21.4	19.0	6.9	5.1	10.2
	Gemini2.5-Pro-Flash	14.7	15.2	**5.1**	11.9	23.5	2.5	20.6	20.3	5.1	4.8	10.5
	Gemini-2.5-Pro-Preview	15.7	14.0	5.1	**10.9**	22.6	2.3	**18.7**	19.2	4.7	4.7	10.3
CR ↑	LLaVA-Video-7B	38.3	37.4	74.9	53.3	15.0	67.7	31.2	43.4	50.4	85.2	53.1
	Qwen2VL-7B	39.8	36.2	75.3	40.2	17.7	73.5	20.1	37.7	58.3	89.1	52.5
	VideoLLaMA3-7B	40.4	37.1	71.0	44.8	19.4	73.3	51.4	46.3	55.3	84.3	54.1
	NVILA-8B	48.6	39.9	77.7	46.4	32.5	77.6	22.2	50.8	57.1	86.2	57.2
	InternVL2.5-8B	46.2	38.2	77.7	46.1	47.1	91.1	69.6	63.2	62.6	91.7	63.3
	Qwen2.5VL-7B	51.6	44.6	84.2	51.2	38.3	91.0	67.2	60.9	67.2	94.3	65.8
	Qwen2.5-VL-72B	57.5	49.7	85.8	52.1	45.8	95.2	84.1	71.5	71.8	94.8	70.9
	GPT-4o	65.4	55.9	88.7	53.7	61.8	95.5	70.7	77.8	67.1	**95.7**	74.1
	Gemini2.5-Pro-Flash	82.5	74.9	92.3	**72.3**	88.0	**98.4**	89.1	93.3	77.0	90.6	84.4
	Gemini-2.5-Pro-Preview	**83.7**	**76.9**	**92.7**	71.3	**88.8**	98.0	**92.1**	**93.7**	**80.9**	92.7	**85.5**

Table 1: The accuracy (AR), inconsistency rate (IR), and coverage rate (CR) of VLM methods on all dimensions, where GPT-4.1 as a TextQA expert. The symbol "↑" indicates that the larger the value, the better; The symbol "↓" indicates that the smaller the value, the better.

substantial improvements in video understanding tasks, with Gemini-2.5-Pro-Preview currently demonstrating the best overall performance.

Gemini-2.5-Pro-Preview excels in generating video descriptions, demonstrating high accuracy (AR), consistency (IR), and comprehensive coverage (CR). This underscores its

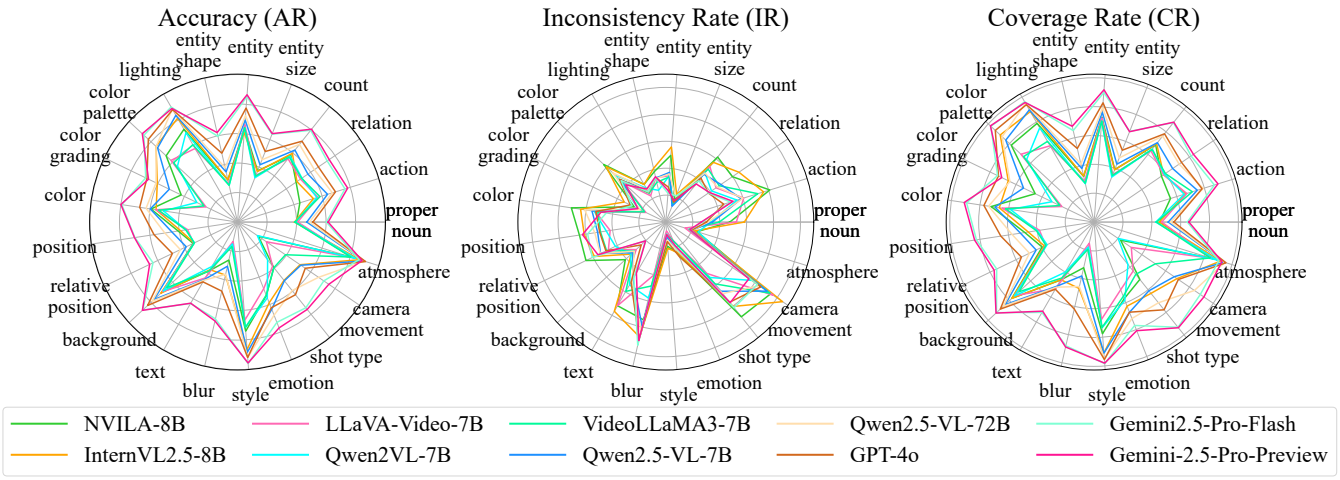

Figure 6: Results of GPT-4.1 captioning evaluation, organized by category.

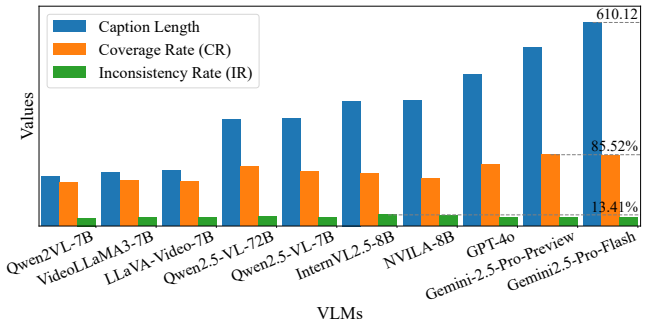

Figure 7: The relationship between CR, IR and caption length, where GPT-4.1 as a TextQA expert. Caption length is the number of words in the caption, not counting special symbols.

superior performance in producing high-quality, detailed, and consistent descriptions compared to other models. VCaps-Bench highlights the shortcomings of existing open-source VLMs, such as the lack of detailed object shapes, sizes, colors, and lighting in descriptions. These deficiencies can be particularly problematic for applications requiring highly detailed and complete captions, such as text-to-video generation models and advanced video analysis tools.

Evaluation Analysis

Gemini surpasses open-source VLMs in every aspect, verifying the advanced video comprehension ability. However, it is crucial to acknowledge that Gemini's captions were utilized in creating QA-pairs, potentially influencing its elevated CR. Nevertheless, Gemini's metrics are valuable as they establish a benchmark for open-source models, emphasizing the disparity in video content understanding. This insight is key in guiding the optimization of open-source VLMs.

Caption length distribution analysis. We also explore the caption lengths to determine if longer captions are associated with a higher evaluation coverage rate (CR), as shown in Fig. 7 . The Fig. 7 compare the performance of various Vision-Language Models (VLMs) in terms of caption length, coverage rate (CR), and inconsistency rate (IR). As shown in Fig. 7, advanced models such as Gemini-2.5-Pro-Flash, Gemini-2.5-Pro-Preview, and GPT-4o generate much longer captions and achieve higher coverage rates compared to earlier models like Qwen2VL-7B and VideoLLaMA3-7B. Despite the increase in caption length, the inconsistency rate remains low and stable across all models.

However, this increase in length also leads to more errors in some open-source models, such as Qwen2VL-7B, VideoLLaMA3-7B, LLaVA-Video-7B, and InternVL2.5-8B. Interestingly, open-source VLMs like Qwen2.5-VL-72B and the closed VLM Gemini-2.5, despite producing longer captions, exhibit lower error rates compared to other VLMs. Additionally, Qwen2.5-VL-7B achieves a higher CR than VILA-8B with shorter captions and a lower IR than LLaVA-Video-7B, which also has shorter captions. This suggests that a deep understanding of video content allows models to generate concise, yet thorough and accurate descriptions.

Conclusion

In this work, we introduce VCapsBench, a new large-scale fine-grained benchmark for evaluating video caption quality. VCapsBench is meticulously designed to support detailed long captions, aiming to advance research and benchmarking in video understanding. The benchmark comprises over 5K videos and more than 100K QA-pairs, assessing 21 critical dimensions of video generation. We have evaluated the caption quality produced by various open-source and closed-source models using this benchmark. The comprehensive analysis highlights the strengths and weaknesses of these models in generating accurate and detailed captions. We believe that VCapsBench will play a crucial role in guiding the optimization of video caption generation, thereby advancing the development of text-to-video models and enhancing the overall understanding of video content.

Acknowledgments

This research is supported by National Science and Technology Major Project (2022ZD0119202), National Science Fund for Distinguished Young Scholars (62125601), National Natural Science Foundation of China (62576031), and State Key Laboratory of Multimedia Information Processing Open Fund (SKLMIP-KF-2025-03).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Bao, F.; Xiang, C.; Yue, G.; He, G.; Zhu, H.; Zheng, K.; Zhao, M.; Liu, S.; Wang, Y.; and Zhu, J. 2024. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Lin, B.; Tang, Z.; et al. 2024a. ShareGPT4Video: Improving Video Understanding and Generation with Better Captions. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chen, T.; Siarohin, A.; Menapace, W.; Deyneka, E.; Chao, H.; Jeon, B. E.; Fang, Y.; Lee, H.; Ren, J.; Yang, M.; and Tulyakov, S. 2024b. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 13320–13331. IEEE.

Cho, J.; Hu, Y.; Baldridge, J. M.; Garg, R.; Anderson, P.; Krishna, R.; Bansal, M.; Pont-Tuset, J.; and Wang, S. 2024. Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation. In *ICLR*.

Farré, M.; Marafioti, A.; Tunstall, L.; Von Werra, L.; and Wolf, T. 2024. FineVideo.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24108–24118.

Grauman, K.; Westbury, A.; Byrne, E.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Liu, M.; Liu, X.; et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18995–19012.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, 7514–7528.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Hu, A.; Chen, S.; Zhang, L.; and Jin, Q. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3171–3185.

Hu, X.; Wang, R.; Fang, Y.; Fu, B.; Cheng, P.; and Yu, G. 2024. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*.

Jiang, M.; Huang, Q.; Zhang, L.; Wang, X.; Zhang, P.; Gan, Z.; Diesner, J.; and Gao, J. 2019. TIGEr: Text-to-Image Grounding for Image Caption Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2141–2152.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. HunyuanVideo: A Systematic Framework For Large Video Generative Models. *arXiv preprint arXiv:2412.03603*.

Lee, H.; Scialom, T.; Yoon, S.; Dernoncourt, F.; and Jung, K. 2021. QACE: Asking Questions to Evaluate an Image Caption. In *2021 Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 4631–4638. Association for Computational Linguistics (ACL).

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Li, D.; Liu, Y.; Wu, H.; Wang, Y.; Shen, Z.; Qu, B.; Niu, X.; Wang, G.; Chen, B.; and Li, J. 2024a. Aria: An Open Multimodal Native Mixture-of-Experts Model. *arXiv preprint arXiv:2410.05993*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, Y.; Chen, X.; Hu, B.; Wang, L.; Shi, H.; and Zhang, M. 2024c. VideoVista: A Versatile Benchmark for Video Understanding and Reasoning. *arXiv preprint arXiv:2406.11303*.

Liu, B.; Akhgari, E.; Visheratin, A.; Kamko, A.; Xu, L.; Shrirao, S.; Lambert, C.; Souza, J.; Doshi, S.; and Li, D. 2024. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*.

Liu, R.; Li, C.; Tang, H.; Ge, Y.; Shan, Y.; and Li, G. 2025a. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, 1–18. Springer.

Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2025b. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4122–4134.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Tan, Z.; Yang, X.; Qin, L.; and Li, H. 2024. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*.

Tian, Y.; Yang, L.; Yang, H.; Gao, Y.; Deng, Y.; Chen, J.; Wang, X.; Yu, Z.; Tao, X.; Wan, P.; et al. 2024. Videotetris: Towards compositional text-to-video generation. *Advances in Neural Information Processing Systems*, 37: 29489–29513.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, S.; Yao, Z.; Wang, R.; Wu, Z.; and Chen, X. 2021. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14050–14059.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Ding, M.; Gu, X.; Huang, S.; Xu, B.; et al. 2025. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22958–22967.

Wang, Y.; Tan, Z.; Wang, J.; Yang, X.; Jin, C.; and Li, H. 2024b. Lift: Leveraging human feedback for text-to-video model alignment. *arXiv preprint arXiv:2412.04814*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S. K.; and Feng, J. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2636–2645.

Yuan, S.; Huang, J.; Xu, Y.; Liu, Y.; Zhang, S.; Shi, Y.; Zhu, R.-J.; Cheng, X.; Luo, J.; and Yuan, L. 2024. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37: 21236–21270.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, S.-X.; Wang, H.; Zhu, X.; Gu, W.; Zhang, T.; Yang, C.; Liu, W.; and Yin, X.-C. 2024a. Video-Language Alignment Pre-training via Spatio-Temporal Graph Transformer. *arXiv e-prints*, arXiv–2407.

Zhang, S.-X.; Yang, C.; Zhu, X.; Zhou, H.; Wang, H.; and Yin, X.-C. 2024b. Inverse-like antagonistic scene text spotting via reading-order estimation and dynamic sampling. *IEEE Transactions on Image Processing*, 33: 825–839.

Zhang, S.-X.; Zhu, X.; Chen, L.; Hou, J.-B.; and Yin, X.-C. 2022. Arbitrary shape text detection via segmentation with probability maps. *IEEE transactions on pattern analysis and machine intelligence*, 45(3): 2736–2750.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024c. Video Instruction Tuning With Synthetic Data. arXiv:2410.02713.

Zhou, H.; Zhu, X.; Zhu, J.; Han, Z.; Zhang, S.-X.; Qin, J.; and Yin, X.-C. 2023. Learning correction filter via degradation-adaptive regression for blind single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12365–12375.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding. *arXiv preprint arXiv:2406.04264*.