OwlCap: Harmonizing Motion-Detail for Video Captioning via HMD-270K and Caption Set Equivalence Reward

Chunlin Zhong[1*], Qiuxia Hou[2*], Zhangjun Zhou[1*], Yanhao Zhang[2+], Shuang Hao[1,3], Haonan Lu[2], He Tang[1✉], Xiang Bai[1✉]

[1]School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China
[2]OPPO AI Center, OPPO Inc., China
[3]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China
{clzhong, hetang, xbai}@hust.edu.cn, {houqiuxia, zhangyanhao}@oppo.com

Abstract

Video captioning aims to generate comprehensive and coherent descriptions of the video content, contributing to the advancement of both video understanding and generation. However, existing methods often suffer from motion-detail imbalance, as models tend to overemphasize one aspect while neglecting the other. This imbalance results in incomplete captions, which in turn leads to a lack of consistency in video understanding and generation. To address this issue, we propose solutions from two aspects: 1) Data aspect: We constructed the Harmonizing Motion-Detail 270K (**HMD-270K**) dataset through a two-stage pipeline: Motion-Detail Fusion (MDF) and Fine-Grained Examination (FGE). 2) Optimization aspect: We introduce the Caption Set Equivalence Reward (**CSER**) based on Group Relative Policy Optimization (GRPO). CSER enhances completeness and accuracy in capturing both motion and details through unit-to-set matching and bidirectional validation. Based on the HMD-270K supervised fine-tuning and GRPO post-training with CSER, we developed **OwlCap**, a powerful video captioning Multi-modal Large Language Model (MLLM) with motion-detail balance. Experimental results demonstrate that OwlCap achieves significant improvements compared to baseline models on two benchmarks: the detail-focused VDC (+4.2 Acc) and the motion-focused DREAM-1K (+4.6 F1).

Code/Dataset — https://github.com/clzhongg/OwlCap

Figure 1: Video captioning encourages capturing both motion and detail. Qualitative (a) and quantitative (b) comparisons show that OwlCap manages to cover both aspects.

Introduction

Video captioning is intended to generate comprehensive and detailed descriptions based on video content. Compared with image captioning, video captioning must depict static details while simultaneously capturing temporal motion specific to video. It has become an important research domain in Multimodal Large Language Models (MLLMs) whose advances significantly impact other fields, including streaming video understanding (Xiong et al. 2025), text-to-video generation (Bar-Tal et al. 2024), and video editing (Chai et al. 2023). With the development of MLLMs, some studies (Chai et al. 2025; Yuan et al. 2025) have focused on improving performance in video captioning across two key dimensions: detail and motion. AuroraCap (Chai et al. 2025) enhances detail extraction capabilities by retraining on task-specific datasets, generating more detailed captions compared to traditional video captioning models. Tarsier (Wang et al. 2024) strengthens motion and event capture in video modalities via large-scale pre-training and 150K human-annotated action-oriented video captions. However, while

*Equal contribution.
+Project Leader.
✉Corresponding author.

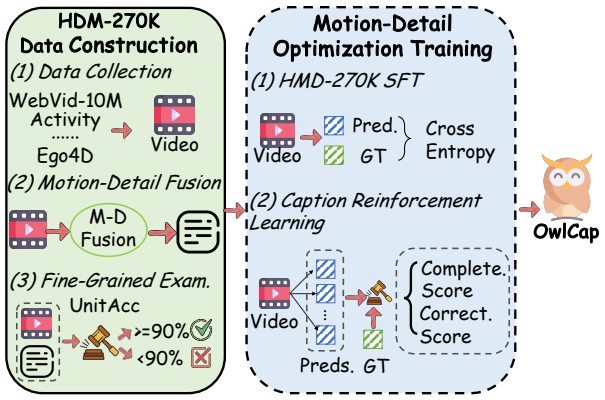

Figure 2: Pipeline for HMD-270K and OwlCap creation.

enhancing one aspect, they often overlook the other, optimizing for either detail characterization or motion capture in isolation, rather than achieving joint optimization of both dimensions: (1) as shown in Figure 1(a), Tarsier typically focuses on outputting motion information in videos, while AuroraCap emphasizes detail presentation, both neglecting the other aspect; (2) as illustrated in Figure 1(b), existing models tend to excel in specific benchmarks: Tarsier performs well on the motion-focused DREAM-1K benchmark (Wang et al. 2024), while AuroraCap excels on the detail-focused Video Detailed Captions (VDC) benchmark (Chai et al. 2025). In summary, existing MLLMs fail to balance motion and detail in video captioning. This imbalance results in incomplete captions, which in turn leads to a lack of consistency in video understanding and generation.

To address this, as shown in Figure 2, we optimize motion-detail from two aspects:

1) Data aspect: We collect videos from various open-source datasets and construct the Harmonizing Motion-Detail 270K (**HMD-270K**) dataset using a two-stage pipeline. First, Motion-Detail Fusion (MDF) leverages distinct MLLMs to separately extract motion and detail information, then integrates them into a fused caption. Second, Fine-Grained Examination (FGE) validates captions by decomposing them into units for individual verification. As shown in Table 1, we calculated words and verbs per second of video in each dataset, and used these statistics to reflect the granularity of detail and motion in the captions. Based on these two statistics, we designed a Motion-Detail Balance (MDB) to measure the balance of the datasets. These data indicate that HMD-270K contains more balanced and comprehensive motion-detail descriptions.

2) Optimization aspect: We first perform Supervised Fine-Tuning (SFT) using the HMD-270K dataset, enabling the model to acquire motion-detail balance capabilities through training on this dataset. Furthermore, we introduce the Caption Set Equivalence Reward (**CSER**) based on the Group Relative Policy Optimization (GRPO) (Shao et al. 2024). Guided by set equivalence theory (Cantor 1874), CSER encourages the model to pursue completeness and correctness of captions through a unit-to-set matching and bidirectional validation strategy, thereby harmonizing the capture of motion and detail information. Compared with VideoCap-R1's Event Score (Meng et al. 2025), which evaluates the single-sided ground-truth event entailed by the predicted caption, CSER adopts a more fine-grained and bidirectional approach to optimize both correctness and completeness of captions.

Based on HMD-270K SFT and CSER, we develop **OwlCap**, a powerful video captioning MLLM that captures both motion and detail information in videos. As shown in Figure 1(b), OwlCap outperforms in both the motion-focused DREAM-1K and the detail-focused VDC benchmarks simultaneously, achieving an accuracy of 46.8% on VDC_Detail and a Recall of 35.3% on DREAM-1K. Furthermore, OwlCap outperforms existing video captioning methods in the Text-to-Video (T2V) generation task. Our key contributions are summarized as follows:

- We introduce HMD-270K, a large-scale dataset comprising 270K video-caption pairs that contain comprehensive motion-detail information, constructed via a designed two-stage pipeline.
- To address the lack of fine-grained modeling in video captioning methods, we introduce CSER, a reward function that employs unit-to-set matching and bidirectional validation to optimize completeness and correctness of generated captions.
- We propose OwlCap, which uses HMD-270K SFT and reinforcement learning with CSER, to enable balance motion-detail caption generation, and our approach outperforms all existing models on mainstream video captioning benchmarks.

Related Work

Video Captioning

Traditional video captions generate a brief text to summarize the main content of a video (Yan et al. 2022; Yang et al. 2023). With the advancement of MLLMs, achieving more comprehensive and coherent video captioning has emerged as a key direction for model optimization (Chen et al. 2024; Yang et al. 2024; Chai et al. 2025). For instance, ShareGPT4Video (Chen et al. 2024) leverages detailed video captions generated by GPT-4V and trains the model using a differential video captioning strategy, enabling it to better describe detailed information. Taesier (Wang et al. 2024) undergoes pre-training on 40 million video-text pairs and further fine-tunes with 150K manually annotated event-focused data, thereby enhancing the model's motion ability. However, these studies fail to effectively balance motion and detail. In contrast, our work addresses this issue by generating large-scale motion-detail balance video-caption pairs through a two-stage pipeline and designing a dedicated caption reward for video captioning.

Reinforcement learning for MLLMs

Recently, numerous studies have begun exploring the application of reinforcement learning to multimodal tasks (Liu et al. 2025; Feng et al. 2025; Li et al. 2025): Visual-RFT (Liu

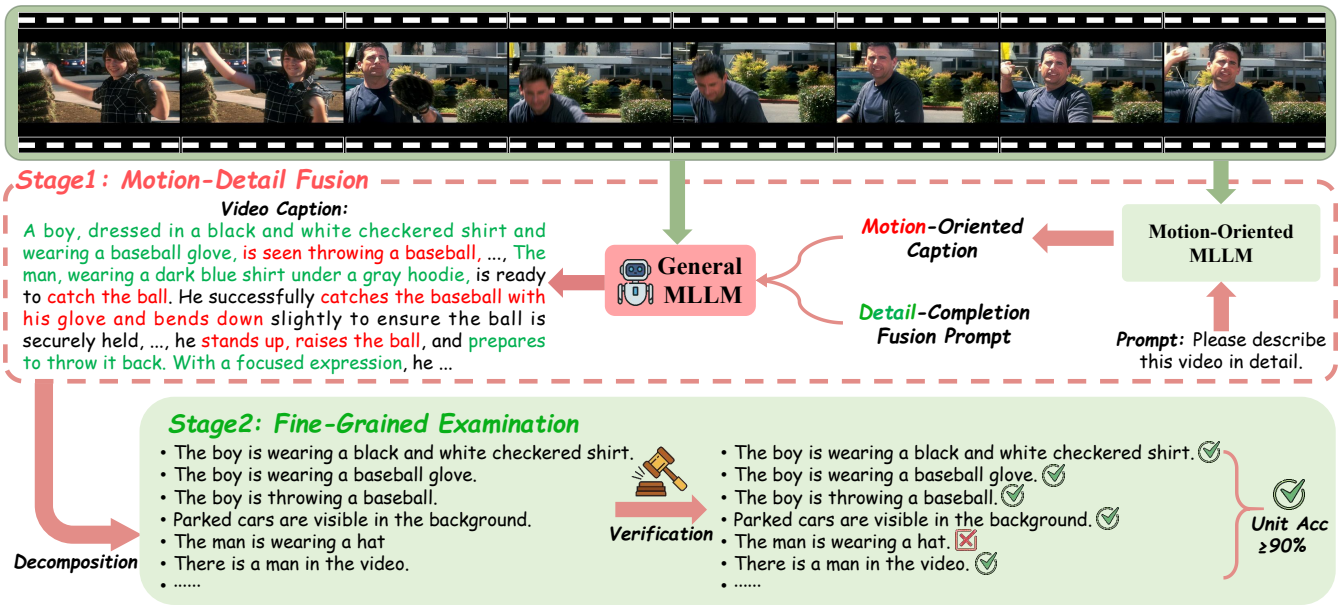

Figure 3: Pipeline for generating high-quality video captions that simultaneously accommodate motion and detail information.

et al. 2025) employs a GRPO-based strategy for various visual perception tasks, improving model performance under limited data. Similar efforts have emerged in the video understanding domain. Video-R1 (Feng et al. 2025) utilizes temporal-based T-GRPO in multiple video tasks, enhancing the model's video understanding capability by integrating GRPO with temporal information. VideoCap-R1 (Meng et al. 2025) attempts to use event coverage as a reward function to improve video caption quality, but its calculation method is overly simplistic, failing to adequately verify the correctness and completeness of captions. Our work further optimizes the model's captioning ability based on GRPO through more fine-grained unit division and more accurate calculation of rewards for correctness and completeness.

HMD-270K Dataset

We introduce a new dataset for video captioning, Harmonizing Motion-Detail (HMD-270K), which contains 270K video-caption pairs capturing both video motions and detailed descriptions. We will first introduce the dataset collection, then illustrate the construction pipeline of HMD-270K, and finally present its statistics.

Dataset Collection

To ensure the diversity of the dataset, we take both dynamic and static information of videos into account during the video collection process. Following Tarsier2-Recap-585K (Yuan et al. 2025), we collected a large pool of videos drawn from the open-source datasets such as WebVid-10M (Bain et al. 2021), ActivityNet (Krishna et al. 2017), and Ego4D (Grauman et al. 2022). These datasets cover multiple domains, including wildlife, movies, cooking, sports, news, and TV programs, providing a solid foundation for understanding various real-world scenarios. **Note that the videos used in the evaluation benchmark are not included in the HMD-270K.**

Dataset Construction Pipeline

As shown in Figure 3, we designed an information fusion and filtering pipeline that leverages open-source video captioning MLLMs to generate motion-detail balanced captions through Motion-Detail Fusion (MDF) and Fine-Grained Examination (FGE) stages. The MDF stage completes the captions, while the FGE stage verifies their accuracy.

Motion-Detail Fusion Stage. The defining characteristic of video captioning lies in two aspects: compared with image captioning, videos contain comprehensive temporal-related motion information specific to video; Furthermore, video captioning requires detailed descriptions of static elements—including object attributes, environmental features, and scene ambiance. To optimize these aspects during dataset construction, as shown in Figure 3, we first input the video into the motion-oriented MLLM Tarsier (Wang et al. 2024). Guided by a prompt designed to focus on temporal motion description, we generate a caption centered on temporal motion. After generation, this motion-focused caption and the original video are jointly fed into the general multimodal large model Qwen2.5-VL-72B (Bai et al. 2025). Through a prompt that guides it to supplement static details, we ultimately generate a video caption that simultaneously contains temporal motion and static details.

Fine-Grained Examination Stage. Compared to other video understanding tasks, video captioning requires describing all video elements. This makes it highly challenging to directly use MLLMs to determine whether a generated caption aligns with the video content. To address this, we design a dedicated "examination" to evaluate the consistency between the generated caption and the video content.

Dataset	Words Per second	Verbs Per second	MDB ↑
Panda-70M	1.5	0.2	0.22
ShareGPT4Video	10.7	1.3	0.54
Vript	13.0	1.4	0.50
Tarsier-585K	5.2	0.8	0.49
HMD-270K	**13.9**	**2.0**	**0.68**

Table 1: Comparison of different video caption datasets. MDB stands for Motion-Detail Balance, its formula and detailed introduction can be found in "Data Statistics".

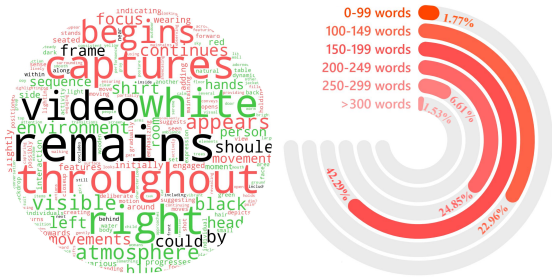

Figure 4: Word cloud (left) and length distribution (right) of captions in HMD-270K.

As illustrated in Figure 3, we decompose the video caption into indivisible units and validate each unit against the video using the Judge Model (InternVL2.5-78B in our pipeline). The motivation behind FGE stems from the fact that this decomposition reduces ambiguity and markedly enhances the transparency and interpretability (Ye et al. 2025). Lastly, we set a 90% unit accuracy (Unit Acc) threshold to filter video-caption pairs and address two challenges: subjective captions (e.g., "exuding a warm atmosphere") causing inconsistent judgments, and ambiguous video targets (e.g., "speaking" vs. "arguing") that cannot be clearly distinguished from frames alone. This threshold provides model fault tolerance.

Data Statistics

The HMD-270K dataset, constructed via a two-stage pipeline, captures both motion and detailed information in videos. Table 1 provides statistics on the number of words and verbs per second for various video captioning datasets, indicating the comprehensiveness of their descriptions of motion and detail. To quantify the balance between motion and detail, we introduced the Motion-Detail Balance (MDB) metric, calculated using the following formula:

$$\text{MDB} = \left(1 - \frac{w - v}{w + v}\right) \cdot \log(w + 1), \tag{1}$$

where w and v denote the number of caption words and verbs per second, respectively. MDB comprehensively accounts for both Motion and Detail in captions. Specifically, $(1 - \frac{w-v}{w+v})$ measures the proportion of motion-related words relative to the total number of words, while $\log(w+1)$ quantifies the level of detail using a logarithmic scale. Figure 4 (left) illustrates the balanced distribution of words related to motion (red color) and detail (green color). Figure 4 (right) depicts the length distribution of captions in HMD-270K, which follows a normal distribution with a mean length of approximately 200 words.

Methodology

To further alleviate the issues of omissions and incorrect outputs in captioning motions and details by MLLMs using the HMD-270K dataset, inspired by reinforcement learning (RL), we designed a Caption Set Equivalence Reward (CSER) based on the GRPO algorithm to improve caption correctness and completeness.

Group Relative Policy Optimization

GRPO (Shao et al. 2024) is a variant of Proximal Policy Optimization (PPO) (Schulman et al. 2017), a policy gradient-based RL algorithm. Unlike PPO, GRPO estimates advantages by comparing candidate answers, without relying on a critic model. Specifically, for a given query q_i, GRPO generates multiple distinct candidate outputs $\{o_1, o_2, o_3, \ldots, o_n\}$ based on the current policy $\pi_{\theta_{\text{old}}}$. These outputs are then evaluated using a predefined reward function to obtain corresponding rewards $r_1, r_2, \ldots, r_n$. The relative quality A_i of these candidate answers is determined by calculating the mean and standard deviation of these rewards:

$$A_i = \frac{r_i - \text{mean}\left(\{r_1, \ldots, r_n\}\right)}{\text{std}\left(\{r_1, \ldots, r_n\}\right)}. \tag{2}$$

GRPO encourages the model to prioritize the responses with higher advantages within the group by updating the policy π_θ using the following clipped surrogate objective:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q,\{o_i\}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Bigg(\min \Bigg(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i, \\ \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i \Bigg) \\ - \beta\, \mathbb{D}_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) \Bigg) \Bigg], \end{aligned} \tag{3}$$

where $\mathbb{E}$ means the expectation operation, G Denotes the number of group, $\text{clip}(\cdot, 1-\epsilon, 1+\epsilon)$ is the clipping function that restricts the input value to the interval $[1-\epsilon, 1+\epsilon]$. $\mathbb{D}_{\text{KL}}$ Stands for Kullback–Leibler Divergence between the new and the reference (old) policies to stabilize updates.

Caption Set Equivalence Reward

The reward function is crucial in reinforcement learning, guiding the model's optimization direction (as shown in Equations (2) and (3)). To address overlooked video elements and erroneous descriptions in video captioning, we developed the Caption Set Equivalence Reward (CSER) functions. CSER is an elegant approach that ensures the correctness and completeness of the predicted captions by leveraging the set equivalence theory (Cantor 1874), which means that for every element in one set, there is exactly one corresponding element in the other set, and vice versa. Specifically, we introduce two components: Correctness Score ensures that units of the predicted caption set match

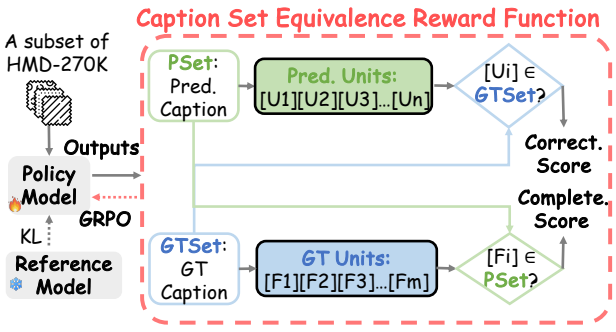

Figure 5: Caption reinforcement fine-tuning with GRPO.

the GT caption set, thereby preventing erroneous outputs. Meanwhile, Completeness Score ensures that units of the GT caption set are fully represented in the predicted caption set, guaranteeing full coverage of GT information. Together, these scores encourage semantic equivalence between the predicted and GT captions through a unit-to-set matching and bidirectional validation strategy, thereby ensuring the model accurately captures all video elements and matches the captions precisely to the video content.

As shown in Figure 5, aligned with the Fine-Grained Examination stage in the HMD-270K construction pipeline, we decompose both the predicted and GT captions into minimal semantic units using the Qwen3-32B large language model. This decomposition follows the principle of breaking down descriptions into irreducible semantic elements. After decomposition, the predicted caption is represented as a sequence of prediction units: $U_1, U_2, ..., U_n$, while the GT caption is transformed into a sequence of GT fact units: $F_1, F_2, ..., F_m$. We then assess the generation quality using two scores based on the concept of set equivalence:

- **Correctness Score**: Measured by the matching accuracy between predicted units and the entire GT Caption. If a single predicted unit fails to match the GT Caption as a whole, it indicates that the unit contains information irrelevant to the video content, thereby reducing the overall accuracy of the generated caption.

- **Completeness Score**: Calculated as the proportion of GT facts covered by the entire predicted caption. If a fact in the GT is not covered by the predicted caption as a whole, it signifies that this key part of the video content is not described, thus reducing the overall completeness of the generated caption.

The specific formulas are as follows:

$$S_{correctness} = \frac{\sum_{i=1}^{n} \mathbb{I}(U_i \in C_{gt})}{n}, \tag{4}$$

$$S_{completeness} = \frac{\sum_{j=1}^{m} \mathbb{I}(F_j \in C_{pred})}{m}, \tag{5}$$

where C_{pred} and C_{gt} denote the predicted caption and the ground truth caption, respectively; $S_{correctness}$ and $S_{completeness}$ denote the Correctness Score and Completeness Score; n and m represent the number of units and facts, respectively. We use the Qwen3-32B model to assess the relevance of units to the GT caption and the coverage of facts by the predicted caption. Notably, we encourage semantic equivalence rather than strict equality (Halmos 1960) between the predicted and GT caption sets, which aligns with RL optimization goals.

Training Processes

We adopted Qwen2.5-VL-7B as the base model, with the training process divided into two phases: Supervised Fine-Tuning (SFT) and reinforcement learning training. In the first phase, we performed SFT on the HMD-270K dataset, enabling the model to balance motion and detail better. In the second phase, we selected 6K samples from the HMD-270K dataset, all of which achieved 100% unit accuracy. These samples were then used as the training set for GRPO. This stage of training aims to optimize the model's correctness and completeness, facilitating the generation of higher-quality video captions. The final reward function for GRPO-based training combines multiple scoring components:

$$\text{Reward} = S_{\text{format}} + S_{\text{correctness}} + S_{\text{completeness}}. \tag{6}$$

The format score S_{format} is to enable the model to output responses in the format we desire.

Experiments

Experiment Setups

Implementation details. During video training, the sampling frame rate is fixed at 2 FPS. In the SFT phase, we adopt the Adam optimizer with a learning rate of 1e-5, using a global batch size of 64 for one epoch of training, which takes approximately 26 hours. For the GRPO phase, following Video-R1 (Feng et al. 2025), we set β = 0.001 for the KL penalty. Each machine utilizes 6 GPUs for training and 2 dedicated GPUs for reward inference, with a global batch size of 24. Training for one epoch in this phase takes about 35 hours. All experiments are implemented on 32 H20 (96GB) GPUs.

Benchmarks. We evaluate our model using two mainstream benchmark datasets: Video Detailed Captions (VDC) (Chai et al. 2025) and DREAM-1K (Wang et al. 2024). The VDC benchmark, comprising over 1,000 videos, rigorously evaluates detailed video caption quality, focusing on characterizing video details. DREAM-1K assesses models' ability to capture fine-grained actions and events, emphasizing accurate motion descriptions. For fair comparison, all benchmark tests and metrics are strictly implemented according to the original code from the papers, and comparative model data is sourced from official leaderboards. We further conduct experiments on the T2V-oriented VidCapBench-AE (Chen et al. 2025) and the fine-grained CaReBench (Xu et al. 2024) to demonstrate generalizability.

Main Results

Result on video captioning. We conducted a comprehensive evaluation of OwlCap across two distinct benchmarks, comparing its performance with that of other general MLLMs and specialized video captioning MLLMs.

Model	Average (Acc / Score)	Detailed (Acc / Score)	Camera (Acc / Score)	Short (Acc / Score)	Background (Acc / Score)	Object (Acc / Score)
Video Caption MLLMs						
ShareGPT4Video-8B (Chen et al. 2024)	36.2/1.9	35.6/1.8	33.3/1.8	39.1/1.9	35.8/1.8	37.1/1.9
Vriptor (Yang et al. 2024)	37.7/2.0	38.5/2.0	37.6/2.0	38.4/2.0	37.1/1.9	37.0/1.9
AuroraCap-7B (Chai et al. 2025)	38.2/2.0	41.3/2.1	43.5/2.3	32.1/1.7	35.9/1.8	39.0/2.0
Tarsier-7B (Wang et al. 2024)	40.2/2.1	38.3/2.1	42.6/2.3	41.7/2.2	36.4/1.9	42.2/2.2
VideoCap-R1-7B (Meng et al. 2025)	43.0/2.3	43.8/2.4	41.7/2.3	35.2/1.9	**47.2/2.5**	47.0/2.5
General MLLM						
VILA-v1.5-8B (Lin et al. 2024)	40.5/2.1	42.8/2.2	39.7/2.1	39.3/2.0	39.8/2.1	40.9/2.1
VideoChat2-7B (Li et al. 2024b)	36.5/1.9	40.5/2.1	31.9/1.7	40.2/2.1	34.9/1.8	34.9/1.8
InternVL-v2-8B (Cai et al. 2024)	33.7/2.0	34.9/1.8	39.1/2.1	33.0/1.7	37.5/1.9	44.2/2.2
LLaVA-OneVision-7B (Li et al. 2024a)	38.8/2.0	41.8/2.2	37.6/2.0	41.6/2.1	34.3/1.8	38.8/2.0
LLaVa-Video-7B (Zhang et al. 2024)	39.0/2.0	35.0/1.8	46.1/2.3	32.8/1.7	37.6/1.9	46.2/2.4
Video-R1-7B (Feng et al. 2025)	43.9/2.3	45.6/2.4	42.7/2.2	44.5/2.3	40.6/2.1	45.9/2.3
Qwen2.5-VL-7B (Bai et al. 2025)	42.7/2.2	43.4/2.3	41.3/2.2	42.2/2.3	41.4/2.1	45.2/2.3
OwlCap-7B	**46.9/2.4**	**46.8/2.4**	**46.9/2.5**	**47.0/2.4**	46.5/2.4	**47.1/2.5**

Table 2: Quantitative comparison with 12 cutting-edge competitors on the VDC benchmark. Best results in **bold**, second-best in underlined. Qwen2.5-VL-7B serves as the baseline. Higher values indicate better performance for all metrics.

Model	Overall	
	F1	Recall
Video Caption MLLMs		
ShareGPT4Video (Chen et al. 2024)	19.5	15.4
AuroraCap-7B (Chai et al. 2025)	20.8	18.1
Tarsier-7B (Wang et al. 2024)	34.6	30.2
VideoCap-R1-7B (Meng et al. 2025)	34.2	34.7
General MLLM		
VILA-v1.5-8B (Lin et al. 2023)	29.9	25.8
VideoChat2-7B (Li et al. 2023)	26.6	23.3
InternVL2-8B (Cai et al. 2024)	26.9	24.7
LLaVa-OneVision (Li et al. 2024a)	31.7	29.4
LLaVA-Video-7B (Zhang et al. 2024)	32.5	28.4
Video-R1-7B (Feng et al. 2025)	33.3	34.5
Qwen2.5-VL-7B (Bai et al. 2025)	30.1	29.7
OwlCap	**34.7**	**35.3**

Table 3: Evaluation results on DREAM-1K benchmark.

As clearly illustrated in Table 2, OwlCap, after undergoing the two-stage training process leveraging the HMD-270K dataset, has achieved a visible improvement in every aspect of the VDC benchmark when compared to the baseline model. Meanwhile, it also outperforms the existing general-purpose MLLMs and video captioning MLLMs, demonstrating OwlCap's advantage in detailed description. Similarly, as presented in Table 3, within the DREAM-1K benchmark, which focuses on motion description, OwlCap has also achieved a significant improvement compared to the Baseline. This consistent performance enhancement further serves to validate the effectiveness of OwlCap.

Validating video captioning models via T2V evaluation. To further validate OwlCap's performance, we conducted additional evaluations using the T2V task, a downstream application of video captioning. Specifically, we collected 200 videos (5–30 seconds long) from Pixabay and used Tarsier, AuroraCap, VideoCap-R1, Qwen2.5-VL-7B, and OwlCap to generate captions. These captions were then fed into HunyuanVideo (Kong et al. 2024) to generate corresponding videos. Subsequently, we compared each generated video with its corresponding reference by uniformly sampling 30 frames from both and then computing the evaluation metrics—SSIM, PSNR, and FID—on these aligned frames. As shown in Table 4, OwlCap demonstrated advantages in the T2V task: videos generated from its captions exhibited notably higher quality than those from competing models.

Model	SSMI ↑	PSNR ↑	FID ↓
Tarsier-7B	0.30	28.45	242.35
AuroraCap-7B	0.27	25.97	261.45
VideoCap-R1-7B	0.30	28.60	240.72
Qwen2.5-VL-7B	0.29	28.02	245.09
OwlCap	**0.33**	**29.32**	**231.60**

Table 4: Comparison of model caption effects on T2V task.

Ablation Studies

Effect of MDF and FGE. To validate the effectiveness of two stages in the HMD-270K construction, we selected 4K videos from the HMD-270K and generated corresponding captions using Tarsier (the motion-oriented model) and Qwen2.5-VL-72B (the detail-completion model) for comparative analysis. Inspired by (Yan et al. 2024), we evaluated the Motion and Detail Scores of captions using GPT-4o with specific instructions. As shown in Table 5, the captions generated after Motion-Detail Fusion achieved higher scores in both motion and detail dimensions. For Unit Acc, we use GPT-4o in place of InternVL2.5-78B as the judge model for evaluation. The initial performance in Unit Acc was unsatisfactory. After optimization through Fine-Grained Examination, Unit Acc improved, along with enhancements in both

	Motion Score	Detail Score	Unit Acc
Tarsier	7.32	6.15	-
Qwen2.5-VL-72B	6.99	6.36	-
MDF	7.80	7.02	78.93
FGE	**7.92**	**7.13**	**96.54**

Table 5: Impact of the MDF and FGE in HMD-270K construction pipeline.

HMD SFT	Reward Com.	Reward Cor.	VDC Acc	VDC Score	DREAM-1K F1	DREAM-1K Recall
			42.7	2.21	30.1	29.7
✓			44.8	2.31	31.9	32.1
	✓		43.5	2.27	32.0	31.7
		✓	43.7	2.26	31.7	31.3
	✓	✓	44.2	2.29	32.7	32.9
✓	✓		45.9	2.34	33.5	33.9
✓		✓	46.2	2.35	33.7	34.0
✓	✓	✓	**46.9**	**2.43**	**34.7**	**35.3**

Table 6: Contribution of HMD-270K (HMD) SFT, Completeness (Com.), and Correctness (Cor.) score rewards.

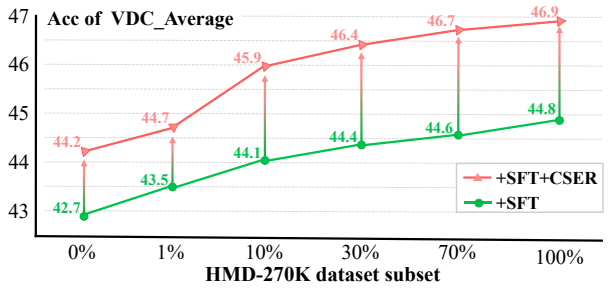

Figure 6: Impact of different ratios in the HMD-270K.

Motion Score and Detail Score.

Effect of HMD-270K and caption score in CSER. Ablation experiments in Table 6 demonstrate how HMD-270K and the correctness and completeness scores in CSER boost video captioning. Key findings include: (i) First-stage SFT training alone offers limited gains; (ii) RL training without prior SFT on HMD-270K also yields minimal improvement; whereas (iii) using the SFT-trained model as the base for RL achieves significant performance enhancements.

Comparison of other Caption reward. To compare CSER with the Event Score in VideoCap-R1 (Meng et al. 2025). We conducted comparative experiments under the same base model. As shown in Table 7, we used the model fine-tuned on HMD-270K via SFT as the baseline, tested the performance of the model on the VDC and DREAM-1K after adding Event Score training. The results indicate that Event Score achieves significant gains on DREAM-1K but only limited improvements on VDC. In contrast, CSER consistently outperforms Event Score across all metrics.

Performance of diverse base models. OwlCap is compatible with multiple Qwen-family models. To investigate the impact of different base models on OwlCap, we present the performance of OwlCap equipped with Qwen2-VL-7B, Qwen2.5-VL-3B, and Qwen2.5-VL-7B in Table 8. It can be clearly seen that, OwlCap consistently outperforms the baselines on both the VDC and DREAM-1K benchmarks, highlighting its stability and effectiveness.

Ablation study about different data sizes. Figure 6 shows the impact of different subsets of HMD-270K with varying proportions on OwlCap. We highlight two key findings: first, the performance improvements become increasingly marginal as the data radio grows; second, SFT training serves as the foundation that enables the subsequent GRPO phase based on CSER to yield further improvements.

Method	Train Time	VDC Acc	VDC Score	DREAM-1K F1	DREAM-1K Recall
Baseline	-	44.8	2.31	31.9	32.1
+ Event Score	+33h	45.2	2.33	32.7	33.2
+ CSER	+35h	**46.9**	**2.43**	**34.7**	**35.3**

Table 7: Comparisons of different caption reward. Event Score is the caption reward proposed in VideoCap-R1.

Model	VDC Acc	VDC Score	DREAM-1K F1	DREAM-1K Recall
Qwen2-VL-7B	39.8	2.1	29.6	26.3
OwlCap	45.2(+5.4)	2.4(+0.3)	34.4(+4.8)	34.7(+8.4)
Qwen2.5-VL-3B	41.1	2.1	29.3	28.9
OwlCap	45.7(+4.6)	2.3(+0.2)	32.5(+3.2)	33.3(+4.4)
Qwen2.5-VL-7B	42.7	2.2	30.1	29.7
OwlCap	46.9(+4.2)	2.4(+0.2)	34.7(+4.6)	35.3(+5.6)

Table 8: OwlCap performance with Qwen family models.

Conclusion

The primary purpose of this paper is to explore how to enhance the quality of video descriptions in video captioning task. We observe that existing methods struggle to simultaneously capture both dynamic actions and fine-grained details in videos. To address this challenge, we propose a video captioning dataset, Harmonizing Motion-Detail 270K (HMD-270K) via a two-stage pipeline. This framework integrates a Motion-Detail Fusion (MDF) and a Fine-Grained Examination (FGE), to synthesize and refine video captions. Furthermore, to fully leverage the potential of HMD-270K, we introduce a Group Relative Policy Optimization (GRPO)-based reinforcement learning strategy, enhanced with a Caption Set Equivalence Reward (CSER), to improve the correctness and completeness of video captioning capabilities for MLLMs. By leveraging HMD-270K for SFT pretraining and GRPO post-training with CSER, we develop OwlCap, a model that effectively captures both motion dynamics and detailed descriptions in videos, overcoming the inherent caption-bias limitation in conventional approaches.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. U25A20403 and Grant No. 62225603), the Natural Science Foundation of Hubei Province of China (No. 2024AFB545), and the Fundamental Research Funds for the Central Universities (Grant No. YCJJ20252416).

References

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*.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.

Bar-Tal, O.; Chefer, H.; Tov, O.; Herrmann, C.; Paiss, R.; Zada, S.; Ephrat, A.; Hur, J.; Liu, G.; Raj, A.; et al. 2024. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Cai, Z.; Cao, M.; Chen, H.; Chen, K.; Chen, K.; Chen, X.; Chen, X.; Chen, Z.; Chen, Z.; Chu, P.; et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Cantor. 1874. Ueber eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chai, W.; Song, E.; Du, Y.; Meng, C.; Madhavan, V.; Bar-Tal, O.; Hwang, J.-N.; Xie, S.; and Manning, C. D. 2025. AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Chen, L.; Wei, X.; Li, J.; Dong, X.; Zhang, P.; Zang, Y.; Chen, Z.; Duan, H.; Tang, Z.; Yuan, L.; et al. 2024. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37: 19472–19495.

Chen, X.; Zhang, Y.; Rao, C.; Guan, Y.; Liu, J.; Zhang, F.; Song, C.; Liu, Q.; Zhang, D.; and Tan, T. 2025. VidCapBench: A Comprehensive Benchmark of Video Captioning for Controllable Text-to-Video Generation. In *In Proceedings of the Findings of the Association for Computational Linguistics*.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wu, J.; Zhang, X.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

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.

Halmos, P. R. 1960. *Naive set theory*. van Nostrand.

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*.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

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, X.; Yan, Z.; Meng, D.; Dong, L.; Zeng, X.; He, Y.; Wang, Y.; Qiao, Y.; Wang, Y.; and Wang, L. 2025. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; and Han, S. 2023. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26689–26699.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Meng, D.; Huang, R.; Dai, Z.; Li, X.; Xu, Y.; Zhang, J.; Huang, Z.; Zhang, M.; Zhang, L.; Liu, Y.; et al. 2025. VideoCap-R1: Enhancing MLLMs for Video Captioning via Structured Thinking. *arXiv preprint arXiv:2506.01725*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Wang, J.; Yuan, L.; Zhang, Y.; and Sun, H. 2024. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.

Xiong, H.; Yang, Z.; Yu, J.; Zhuge, Y.; Zhang, L.; Zhu, J.; and Lu, H. 2025. Streaming Video Understanding and Multi-round Interaction with Memory-enhanced Knowledge. In *The Thirteenth International Conference on Learning Representations*.

Xu, Y.; Li, X.; Yang, Y.; Meng, D.; Huang, R.; and Wang, L. 2024. CaReBench: A Fine-Grained Benchmark for Video Captioning and Retrieval. *arXiv preprint arXiv:2501.00513*.

Yan, S.; Bai, M.; Chen, W.; Zhou, X.; Huang, Q.; and Li, L. E. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, 37–53. Springer.

Yan, S.; Zhu, T.; Wang, Z.; Cao, Y.; Zhang, M.; Ghosh, S.; Wu, Y.; and Yu, J. 2022. VideoCoCa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*.

Yang, A.; Nagrani, A.; Seo, P. H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; and Schmid, C. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10714–10726.

Yang, D.; Huang, S.; Lu, C.; Han, X.; Zhang, H.; Gao, Y.; Hu, Y.; and Zhao, H. 2024. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37: 57240–57261.

Ye, Q.; Zeng, X.; Li, F.; Li, C.; and Fan, H. 2025. Painting with Words: Elevating Detailed Image Captioning with Benchmark and Alignment Learning. In *The Thirteenth International Conference on Learning Representations*.

Yuan, L.; Wang, J.; Sun, H.; Zhang, Y.; and Lin, Y. 2025. Tarsier2: Advancing Large Vision-Language Models from Detailed Video Description to Comprehensive Video Understanding. *arXiv preprint arXiv:2501.07888*.

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