Large Multi-modal Model for Video Captioning

Wenhao Chai

A thesis
submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

University of Washington

2025

Committee:

Jenq-Neng Hwang

Chengcheng Zhu

Program Authorized to Offer Degree:
Electrical & Computer Engineering

University of Washington

Abstract

Large Multi-modal Model for Video Captioning

Wenhao Chai

Chair of the Supervisory Committee:
Jenq-Neng Hwang
Electrical & Computer Engineering

Video detailed captioning is a key task which aims to generate comprehensive and coherent textual descriptions of video content, benefiting both video understanding and generation. In this paper, we propose AuroraCap, a video captioner based on a large multimodal model. We follow the simplest architecture design without additional parameters for temporal modeling. To address the overhead caused by lengthy video sequences, we implement the token merging strategy, reducing the number of input visual tokens. Surprisingly, we found that this strategy results in little performance loss. AuroraCap shows superior performance on various video and image captioning benchmarks, for example, obtaining a CIDEr of 88.9 on Flickr30k, beating GPT-4V (55.3) and Gemini-1.5 Pro (82.2). However, existing video caption benchmarks only include simple descriptions, consisting of a few dozen words, which limits research in this field. Therefore, we develop VDC, a video detailed captioning benchmark with over one thousand carefully annotated structured captions. In addition, we propose a new LLM-assisted metric VDCscore for bettering evaluation, which adopts a divide-and-conquer strategy to transform long caption evaluation into multiple short question-answer pairs. With the help of human Elo ranking, our experiments show that this benchmark better correlates with human judgments of video detailed captioning quality.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Number Page

LIST OF TABLES

GLOSSARY

VIDEO DETAILED CAPTIONING: A task that aims to generate comprehensive and coherent textual descriptions of video content.

LLM: Large Language Model, a type of artificial intelligence model that uses deep learning techniques to understand and generate human-like text.

LARGE MULTIMODAL MODEL: A model that processes and understands multiple types of data, such as text and images, simultaneously.

AURORACAP: A video captioner based on a large multimodal model that follows a simple architecture design without additional parameters for temporal modeling.

VDC: A video detailed captioning benchmark with over one thousand carefully annotated structured captions, developed to facilitate research in video captioning.

VDCSCORE: A new LLM-assisted metric designed to evaluate long captions by breaking down the assessment into multiple short question-answer pairs, aiming to better correlate with human judgments.

TOKEN MERGING STRATEGY: A technique used to reduce the number of input visual tokens by combining tokens, thereby decreasing computational overhead without significantly affecting performance.

ELO RANKING: A method for calculating the relative skill levels of players in competitor-versus-competitor games, adapted here for ranking the quality of generated captions based on human comparisons.

ACKNOWLEDGMENTS

I would like to thank my advisor, Jenq-Neng Hwang, for his guidance and support throughout my research. I would also like to thank my labmates for their help and support.

Chapter 1

INTRODUCTION

The task of video detailed captioning involves generating comprehensive and coherent textual descriptions of video content, capturing not only the primary actions and objects but also intricate details, contextual nuances, and temporal dynamics. It has emerged as a critical area of research in computer vision and natural language processing, with significant implications for the fields of robotics [163, 37], ego-centric perception [47, 48], embodied agents [177, 179, 180, 34, 35, 178], and video editing [14] and generation [7]. The challenges of video detailed captioning compared to past problems include the limited detailed caption data for training and evaluation, and also the lack of a good evaluation metric.

Before the emergence of Large Language Models (LLMs), previous models could only generate very short and rough descriptions of videos [143, 153, 158, 162]. Although these models have been trained on web-scale video-text datasets (*e.g.,* HowTo100M [101] and VideoCC3M [105]), their capabilities remain limited due to their smaller scale and the lack of world knowledge possessed by LLMs. Recently, researchers start to build more powerful large multimodal models (LMMs) upon pretrained LLMs (*e.g.,* LLaVA [92], InstructBlip [32], InternVL [27]). These models typically use intermediate components (*e.g.,* Q-Former [73] or an MLP) to connect the pre-trained vision transformer (ViT) [36] and the LLM. Expanding from image-based LMMs to video-based LMMs is a natural progression, as videos can be viewed as sequences of frames. While most LMMs start with loading the pre-trained weights from image models and are further fine-tuned with video-text data, we find that LLaVA-like models can be easily adapted to a video one without any additional parameters but only with high-quality video-text instruction data for fine-tuning.

However, naive treatment of videos as a series of image frames can result in significant computational overhead and may cause a generalization of the length problem [144]. To address these concerns, more specifically, to reduce the number of visual tokens, Video-

LLaMA [82] adapts the video Q-former, MovieChat [127] uses a memory bank, LLaMA-VID [83] simply uses global pooling, and FastV [20] drops visual tokens by attention rank within LLM layers. In this paper, we present AuroraCap [15], adapting a simple yet efficient method called Token Merging [11], which is proved to be effective in image and video classification and editing tasks [81]. To be specific, we gradually combine similar tokens in a transformer layer using a bipartite soft matching algorithm to reduce the number of visual tokens. Following this pattern, our experiments show that we can use only 10% to 20% visual tokens compared to the original tokens generated by ViT with a marginal performance drop in various benchmarks. With this technique, it is easier to support higher-resolution and longer video sequence inputs for training and inference.

We present results on several widely used benchmarks, but find that existing video understanding benchmarks are either question-answer-based [127, 17, 12, 154, 150, 42, 147], which cannot demonstrate detailed descriptive abilities, or they provide descriptions that are too short, with only a few words [154, 12] as shown in Table 1.1. Some large-scale datasets focus on specific domains such as ego-centric [48] or contain low-quality videos and annotations [5]. Therefore, we construct the VDC (**V**ideo **D**etailed **C**aptions) benchmark, which contains over one thousand high-quality video-caption pairs spanning a wide range of categories, and the resulting captions encompass rich world knowledge, object attributes, camera movements, and crucially, detailed and precise temporal descriptions of events. We utilize `GPT-4o` as our recaption assistant with a hierarchical prompt design. To preserve as much information as possible from the videos and maintain temporal consistency, we implement a dense-frame extraction strategy. Using the dense frames as input, despite description of the whole video, we also generate high-quality captions from different aspects, including objective facts, backgrounds, camera angles and movements. Manual quality inspection is employed to ensure the quality of the video captions. While existing video-caption datasets [23, 61, 71] offer structured captions, VDC is the first benchmark focused on detailed video captioning, providing significantly longer and more detailed captions than others as shown in the Table 1.1 and Section B.

We also introduce a novel evaluation metric specifically designed for detailed captioning task. Traditional metrics like METEOR [6], CIDEr [140], and BLEU [109], designed for

Table 1.1: **Benchmark comparison** for video captioning task. Ave. Length indicates the average number of words per caption. Compared to the existing benchmarks, VDC has the longest captions.

Dataset	Theme	# Video	# Clip	# Caption	# Word	# Vocab.	Ave. Length
MSVD [17]	Open	1,970	1,970	70,028	607,339	13,010	8.67
MSR-VTT [154]		7,180	10,000	200,000	1,856,523	29,316	9.28
ActivityNet [66]		20,000	100,000	100,000	1,340,000	15,564	13.40
S-MiT [104]		515,912	515,912	515,912	5,618,064	50,570	10.89
M-VAD [139]	Movie	92	48,986	55,905	519,933	18,269	9.30
MPII-MD [119]		94	68,337	68,375	653,467	24,549	9.56
Youcook2 [182]	Cooking	2,000	15,400	15,400	121,418	2,583	7.88
Charades [125]	Human	9,848	10,000	27,380	607,339	13,000	22.18
VATEX [146]		41,300	41,300	413,000	4994,768	44,103	12.09
VDC (ours)	Open	1,027	1,027	1,027	515,441	20,419	500.91

machine translation or short captions, fail to evaluate detailed captions which contain rich semantic information. On the other side, an LLM-based evaluation metric is commonly used in visual question answering benchmarks [97, 147], especially for those generated by VLMs [128, 127]. However, we observe that the LLM-based evaluation metric still struggles to differentiate the quality of detailed captions and tends to give lower scores. To address these challenges, we propose VDCSCORE, a novel captioning evaluation metric that leverages the reliability of large language models (LLMs) by evaluating short visual question-answer pairs. We first decompose the ground-truth caption into a set of concise question-answer pairs using the LLM, then generate corresponding responses from the predicted caption. Finally, the LLM is used to assess the accuracy of each response to provide an overall score. In particular, our paper makes the following contributions:

- In Chapter 2, we illustrate how we can reduce the number of tokens used for image or video input before injecting into LLM with marginal performance drop. Using these insights, we propose AURORACAP, which is shown to be the state-of-the-art video captioning model.
- In Chapter 3, we present VDC, the first benchmark for detailed video captioning, featuring over one thousand videos with significantly longer and more detailed captions. We comprehensively evaluate proprietary and open-source models using our proposed VDCSCORE metric.

Chapter 2

AURORACAP: A VIDEO DETAILED CAPTIONING BASELINE

2.1 Architecture

Large multimodal model. To effectively leverage the capabilities of both the pre-trained LLM and visual model, which is typically CLIP [114] or DINO [107], LLaVA adapt a simple multi-layer perceptron (MLP) as projection layer to connect each patch tokens of image features into the word embedding space. The original LLaVA model is trained by a two-stage instruction-tuning procedure, which first pretraining projection layer for feature alignment and then finetuning end-to-end while freeze the visual encoder. Recent works like Prismatic VLMs [62] and Idefics2 [68] further explore the design space of LLaVA architecture. We adapt some conclusion from those works for training the model.

Token merging. To increase the throughput of existing ViT models, Token Merging [11] is proposed to gradually combines similar tokens in a transformer to reduce the number of tokens passing through ViT models. Token Merging has been proven to be effective on image and video classification tasks even without the need for training. Token Merging is applied between the attention and MLP within each transformer block as:

1. Alternatively partition the tokens into two sets $\mathbb{A}$ and $\mathbb{B}$ of roughly equal size.

2. For each token in set $\mathbb{A}$, calculate the token similarity with each token in set $\mathbb{B}$ based on cosine similarity of the *Key* features in attention block.

3. Use bipartite soft matching and then select the most similar r pairs.

4. Merge the tokens using weighted average, record the token size.

5. Concatenate the two sets $\mathbb{A}$ and $\mathbb{B}$ back together again.

Once the tokens have been merged, they actually carry the features for more than one input patch. Therefore, the merged tokens will have less effect in softmax attention as

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{\mathbf{d}}} + \log \mathbf{s}\right) \tag{2.1}$$

where $\mathbf{s}$ is the number of patches the token represents after token merging. We conduct frame-wise token merging in AuroraCap, the visualization of token merging can be found in Appendix C.

2.2 Training Recipe

Building upon the exploration in works like Prismatic VLMs [62], Idefics2 [68], and InternVL [27], we further adopt a three-stage training strategy, which can be noted as Pretraining stage, Vision stage and Language stage. The training data used in each stage are shown in Table E.1, Table E.2 and Table E.3. More training details including hyper-parameters selection and data preprocessing operation can be found in Appendix E.

Pretraining stage. Similar to LLaVA, we first align visual features from the vision encoder ViT with the word embedding space of LLMs. To achieve this, we freeze the pretrained ViT and LLM, training solely the vision-language connector. Consistent with LLaVA-1.5 [70], we employ a two-layer MLP as the projection layer and pretrain on 1.3M image-caption pairs. To optimize performance, we explore various combinations of the pre-trained ViT and LLM in Appendix D.

Vision stage. Unlike LLaVA, we next unfreeze the pretrained ViT while freezing the LLM during vision stage and train with the public data among various computer vision tasks (*e.g.,* captioning, object identification, classification, reasoning, VQA, and *etc.*) to get better generalization [55]. The motivation for doing this is that CLIP ViT usually performs poorly in aspects such as Orientation and Direction, Positional and Relational Context, Quantity and Count [138]. However, since the most of the collected datasets lack high-quality and detailed corresponding language descriptions, the labels often consist of only a few words or a short phrase when converted to text. Therefore, unfreezing the language model at this stage is risky, as it may lead to a degradation in the performance of the language model.

Table 2.1: Comparison of AuroraCap with LLM-based SoTA methods on image captioning benchmarks under zero-shot setting. The number in the upper right corner indicates the number of shots. Scores with **bold** indicate the best performance under zero-shot setting.

Model	Flickr (31,784)					NoCaps (4,500)					COCO-Cap (5,000)				
	C	B@1	B@4	M	R	C	B@1	B@4	M	R	C	B@1	B@4	M	R
LLaVA-1.5-7B [90]	74.9	71.7	28.4	26.1	52.8	105.5	82.6	40.2	30.3	59.4	110.3	73.0	29.7	29.2	55.5
LLaVA-1.5-13B	79.4	73.6	30.2	26.6	53.9	109.2	84.2	42.4	**30.6**	60.3	115.6	74.6	31.5	**29.4**	56.5
LLaVA-1.6-7B [91]	68.4	69.6	26.6	23.2	50.3	88.4	73.8	34.8	25.9	54.6	99.9	67.7	28.4	25.5	52.4
LLaVA-1.6-13B	66.6	65.2	24.2	22.2	48.8	88.1	68.7	34.0	25.4	54.9	101.8	62.2	27.5	24.6	52.1
MiniCPM-V-3B [52]	66.8	68.0	25.1	27.2	51.0	89.9	79.1	33.2	29.7	55.8	94.2	69.8	23.9	28.3	52.3
DeCap [77]	56.7	–	21.2	21.8	–	42.7	–	–	–	–	91.2	–	24.7	25.0	–
Flamingo-80B [2]	67.2	–	–	–	–	–	–	–	–	–	84.3	–	–	–	–
Chameleon-34B [132]	74.7^2	–	–	–	–	–	–	–	–	–	120.2^2	–	–	–	–
GPT-4V	55.3^8	–	–	–	–	–	–	–	–	–	78.5^8	–	–	–	–
Gemini-1.5 Pro	82.2^4	–	–	–	–	–	–	–	–	–	99.8^2	–	–	–	–
AuroraCap-7B	**88.9**	**75.6**	**32.8**	**26.7**	**55.4**	**111.4**	**85.6**	**44.4**	29.9	**60.6**	**120.8**	**78.0**	**35.3**	28.6	**57.2**

Language stage. Finally, we conduct end-to-end training, which means all the components are trainable, with the most high-quality public data during language stage training. We mix all the data, including images and videos, captions and instructions, into each mini-batch for training. To improve video captioning performance, we duplicate the video captioning datasets twice. We remove all the video training data for training a image-based AuroraCap as well for image captioning task.

2.3 Evaluation

In this section we evaluate AuroraCap on various tasks including image captioning, video captioning, and video question answering. Appendix F show detailed evaluation settings.

Image Captioning. We evaluate AuroraCap using CIDEr (C), BELU-4 (B@4), BELU-1 (B@1), METEOR (M), and ROUGE-L (R) metric on Flickr [113], NoCaps [1], and COCO-Cap [88] benchmarks and compare it with LLM-based state-of-the-art methods as shown in Table 2.1. We show the performance of image based AuroraCap under zero-shot settings.

Table 2.2: Comparison of AuroraCap with SoTA methods on existing video captioning benchmarks under zero-shot setting.

Model	MSR-VTT (1,000)					VATEX (1,000)				
	C	B@1	B@4	M	R	C	B@1	B@4	M	R
ZeroCap [134]	9.6	–	2.9	16.3	35.4	–	–	–	–	–
DeCap [77]	18.6	–	14.7	20.4	–	18.7	–	13.1	15.3	–
PaLI-3 [25]	21.3	–	–	–	–	–	–	–	–	–
Ma *et al.* [96]	22.1	–	3.5	17.3	28.7	23.9	–	2.8	14.1	23.5
LLaVA-7B [92]	16.9	–	–	–	–	–	–	–	–	–
Video-LLaMA [169]	2.3	–	4.9	16.8	–	3.8	–	4.3	16.3	21.8
AuroraCap-7B	**33.1**	**58.6**	**21.0**	**23.9**	**49.5**	**33.8**	**57.1**	**18.4**	**19.0**	**40.8**

Notice that these benchmarks all contain short captions consisting of a single sentence, so they only partially reflect the model's performance. The performance mentioned in the rest of this paper refers to video-based AuroraCap.

Video Captioning. Although the current video captioning benchmarks are only contains one-sentence captions, to compare with prior work, we similarly evaluate on these benchmarks. We evaluate AuroraCap using CIDer (C), BELU-4 (B@4), BELU-1 (B@1), METEOR (M), and ROUGE-L (R) metric on MSR-VTT [154], VATEX [146] and compare it with other methods as shown in Table 2.2.

Video Question Answering. We evaluate AuroraCap on MSVD-QA [152], ActivityNet-QA [166], MSRVTT-QA [152], and iVQA [160] for video question answering tasks as shown in Table 2.3. Although AuroraCap is primarily a captioning model, it achieves competitive performance in some VQA datasets (ANet, iVQA). For others (MSVD, MSR-VTT) performance is more modest, but still not bad. In some failure cases observed in the model, we found that prompting the model to generate a comprehensive caption for the

Table 2.3: Comparison of AuroraCap with SoTA methods on video question answering and classification benchmarks under zero-shot setting. The pretrained llm size is 7B for all the listed models.

Model	ANet		MSVD		MSR-VTT		iVQA
	Acc	Score	Acc	Score	Acc	Score	Acc
Just Ask [160]	–	–	–	–	–	–	12.2
FrozenBiLM [161]	24.7	–	32.2	–	16.8	–	26.8
Video-LLaMA [169]	12.4	1.1	51.6	2.5	29.6	1.8	–
VideoChat [75]	26.5	2.2	56.3	2.8	45.0	2.5	–
Video-ChatGPT [97]	35.2	2.7	64.9	3.3	49.3	2.8	–
LLaMA-VID [83]	47.4	3.3	69.7	3.7	57.7	3.2	–
Video-LLaVA [85]	45.3	3.3	70.7	3.9	59.2	3.5	–
FreeVA [149]	51.2	3.5	73.8	**4.1**	**60.0**	**3.5**	–
LLaVA-NeXT-Video [91]	53.5	3.2	–	–	–	–	–
MovieChat [127]	45.7	3.4	75.2	3.8	52.7	2.6	–
MovieChat+ [128]	48.1	3.4	**76.5**	3.9	53.9	2.7	–
AuroraCap-7B	**61.8**	**3.8**	62.6	3.6	43.5	2.9	**55.2**

video input can lead to outputs that include the correct answer. This phenomenon may be attributed to a disruption in the model's instruction-following capabilities during the visual-based training period. We regard this as a promising avenue for future research.

Chapter 3

VDC: A VIDEO DETAILED CAPTIONING BENCHMARK

3.1 Benchmark Dataset Curation

3.1.1 Video Collection and Processing

To ensure the reliability of the benchmark, it is crucial to maintain high video quality, balanced data distribution, and content complexity. Panda-70M [24] offers a high-resolution, open-domain YouTube video dataset with diverse one-minute clips across wildlife, cooking, sports, news, TV shows, gaming, and 3D rendering, ideal for studying complex real-world scenarios. Additionally, a large volume of aesthetically appealing videos from user-uploaded platforms like Mixkit [103], Pixabay [112], and Pexels [110] provides scenic views and visually pleasing human activities with minimal transitions and simpler events. Ego4D [47] complements the video source by focusing on ego-centric human activities and auto-driving scenarios, ensuring comprehensive coverage of real-world scenes. To mitigate content homogeneity among these candidate videos and maintain diversity in the final dataset, inspired by ShareGPT4Video [23], we building VDC upon those various video sources. Note that the videos used in VDC construction are not included in the training data of AuroraCap. To ensure balanced data distribution, we allocate equal proportions of videos from Panda-70M [24], Ego4D [47], Mixkit [103], Pixabay [112], and Pexels [110]. We first split the video into clips and apply dense frame extraction.

3.1.2 Structured Detailed Captions Construction Pipeline

We believe that a comprehensive detailed video caption benchmark should encompass various aspects, including main objects, camera movements, and background. However, most existing benchmarks [71, 23] provide only a single caption for the entire video with less structured details. Therefore, we develop a structured detailed captions construction pipeline to generate extra detailed descriptions from various perspectives, significantly extending the length and

Table 3.1: The video source distribution of the proposed VDC benchamrk including diverse settings such as natural landscapes, human activities, and animal activities. Videos from different sources have a similar proportion in VDC, reducing the data bias.

Video Source	# Sample	Proportion	Duration (sec.)	Ave. Length (sec.)	Ave. # Keyframe
Panda-70M [24]	229	22.25%	5,714	24.95	7.18
Ego4D [47]	196	19.05%	10,935	55.79	19.46
Mixkit [103]	197	19.14%	3,261	16.55	6.58
Pixabay [112]	199	19.34%	4,748	23.86	8.99
Pexels [110]	208	20.21%	4,343	20.88	7.99
Total	1,027	–	29,001	28.18	10.43

enhancing the richness compared to previous benchmarks. Following [61], the structured captions in VDC encompass not only **short** and **detailed captions** but also three additional categories: (1) **main object caption**, offering a comprehensive analysis of the primary subjects' actions, attributes, interactions, and movements across frames, including variations in posture, expression, and speed; (2) **background caption**, providing detailed descriptions of the background, such as objects, location, weather, time, and dynamic elements; and (3) **camera caption**, which details the camera work, including shot types, angles, movements, transitions, and special effects.

To generate detailed, fine-grained, and accurate captions, we leverage `GPT-4o` to produce video descriptions. We utilize the dense video frames to obtain captions. We observed that generating all captions in a single conversation round often introduces hallucinations in the detailed captions. To address this, we design a hierarchical prompt strategy to efficiently obtain accurate structured captions and detailed captions in two conversation rounds: (1) structured captions generation and (2) detailed captions integration. In the first round, the prompt briefly introduces the differences between structured captions and uses the dense video frames as input to generate the short caption, main object caption, background caption,

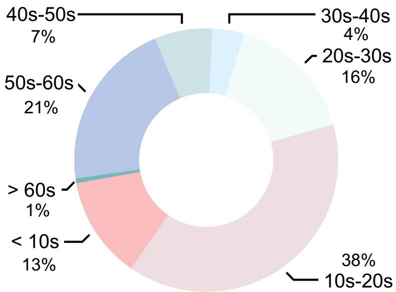


Figure 3.1: Video length statistics.

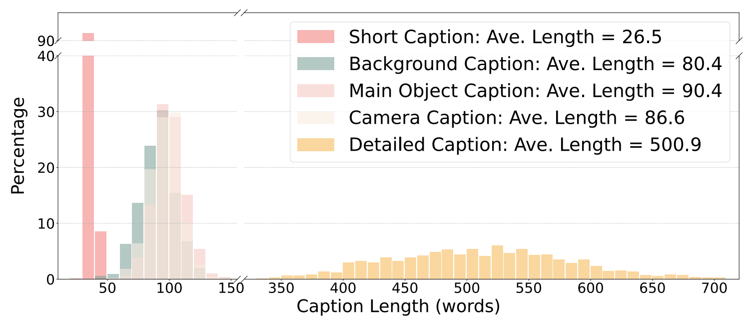


Figure 3.2: Distribution of structured caption length.

camera caption, and the detailed caption. In the second round, the generated captions serve as the reference. The second-round prompt guides `GPT-4o` to enhance the detailed caption based on the initial captions, ensuring consistency without introducing new entities or relations, and producing a vivid, engaging, and informative description. The whole prompt template for the structured detailed captions construction pipeline can be found in Appendix H. Finally, we conduct a manual review to correct captions with hallucinations and supplement omitted visual elements. The refined detail structured captions are then used as the ground truth for evaluation.

3.1.3 Comparison on Numerical Statistics

Based on the hierarchical scheme, VDC can capture a rich variety of details of the video and reduce hallucinations. The visual representation in Figure 3.1 demonstrates the video duration distribution of VDC. Over 87% of the videos exhibit a duration ranging from 10K to 12K frames, while 1% of videos extending beyond 60 seconds. Only 13% of videos have duration less than 10 seconds. As illustrated in Table 1.1, the average length of detailed descriptions in VDC is significantly longer than in previous benchmarks. Figure 3.2 shows the length distribution of structured captions in VDC, with detailed captions averaging over 500 words. Appendix J present more statistics.

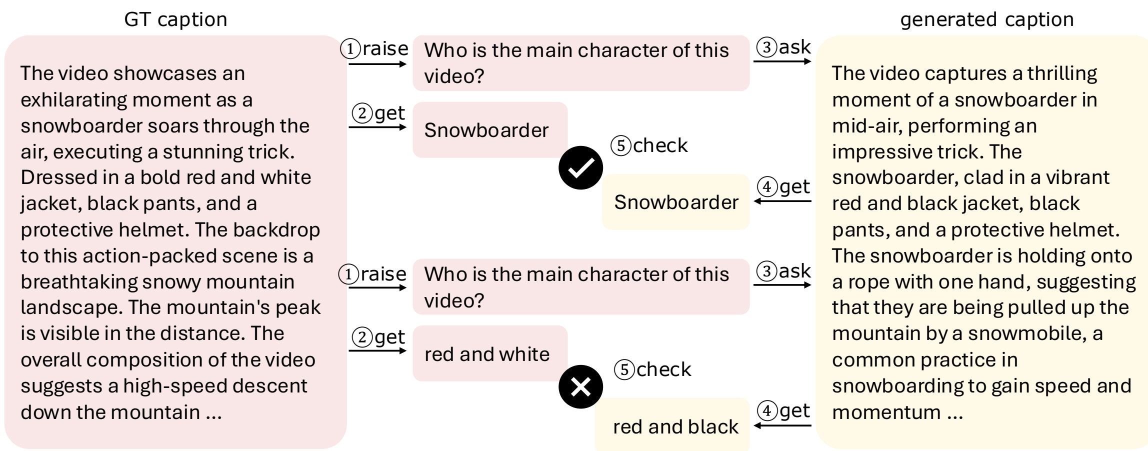


Figure 3.3: Evaluation pipeline with VDCSCORE. Like when humans take reading comprehension tests, we transform the matching between two paragraphs into a set of question-answer pairings. We first generate some question-answer pairs based on the ground truth captions, then derive corresponding answers one by one from the generated captions, and finally perform matching. The process is automatically evaluated with the LLM involvement in each step.

3.2 Evaluation Metric Design and Leaderboard

3.2.1 VDCSCORE*: Evaluating Detailed Captions with LLMs*

Evaluating video captions requires not only assessing the quality of the captions but also flexibly evaluating the alignment between the video and the caption. While metrics such as BLEU [109], CIDEr [140], and ROUGE-L [86] have been employed for caption evaluation, these metrics are predominantly designed for short captions and rely heavily on word-level frequency-based alignment. Given the advanced semantic understanding capabilities of large language models (LLMs), Video-ChatGPT [75] proposes using LLM as an evaluation assistant to directly judge the correctness of the whole predicted captions and assign scores. However, as demonstrated in Table 3.2, our experiments indicate that when dealing with detailed captions, the direct application of LLM struggles to accurately distinguish the correctness of various predicted captions, fails to effectively evaluate the precision of detailed descriptions, and exhibits a tendency to assign disproportionately lower scores. Therefore, we introduce VDCScore, a novel quantitative metric that utilizes LLMs to evaluate the similarity between predicted and ground-truth detailed captions through a divide-and-conquer approach.

The core idea of VDCScore is to decompose long detailed captions into multiple short question-answering pairs, avergae the evaluation of each pair as the final result. We elaborate the design of VDCScore in the following parts: (1) ground-truth question-answer pairs extraction, (2) responsed answers generation and (3) answers matching. As illustrated in Figure 3.3, we first employ `GPT-4o` to generate question-answer pairs from the detailed ground-truth caption. To ensure that the generated question-answer pairs capture as much information as possible from the original caption and facilitate accurate evaluation by LLMs, we constrain the number of generated pairs and impose specific guidelines: the questions must be open-ended, and the answers should be concise, and directly relevant to the questions. For a fair comparison and to mitigate potential variability arising from generating different question-answer pairs for the same caption, we pre-generate a standardized set of question-answer pairs for all captions in VDC, as depicted in Figure 3.3. The used prompt templates used along with additional examples, are provided in Appendix I.

VDCScore subsequently analyzes the predicted captions by leveraging ground-truth question-answer pairs. We prompt `GPT-4o` to read the detailed predicted captions and generate answers based solely on these captions. To mitigate biases arising from discrepancies in the length between ground-truth and predicted answers, we also impose constraints ensuring that responses are limited to concise sentences or phrases. Consequently, for each pair of ground-truth and predicted captions, we obtain a set of `<question, correct answer, predicted answer>` triplets. Following Video-ChatGPT [75], we then ask `GPT-4o` to output two scores for each triplet: one for answer correctness and another for answer quality. The final accuracy and score are calculated by averaging the correctness score and quality score respectively. When using two same captions as input, VDCScore returns an accuracy of 100%, demonstrating the feasibility and reliability.

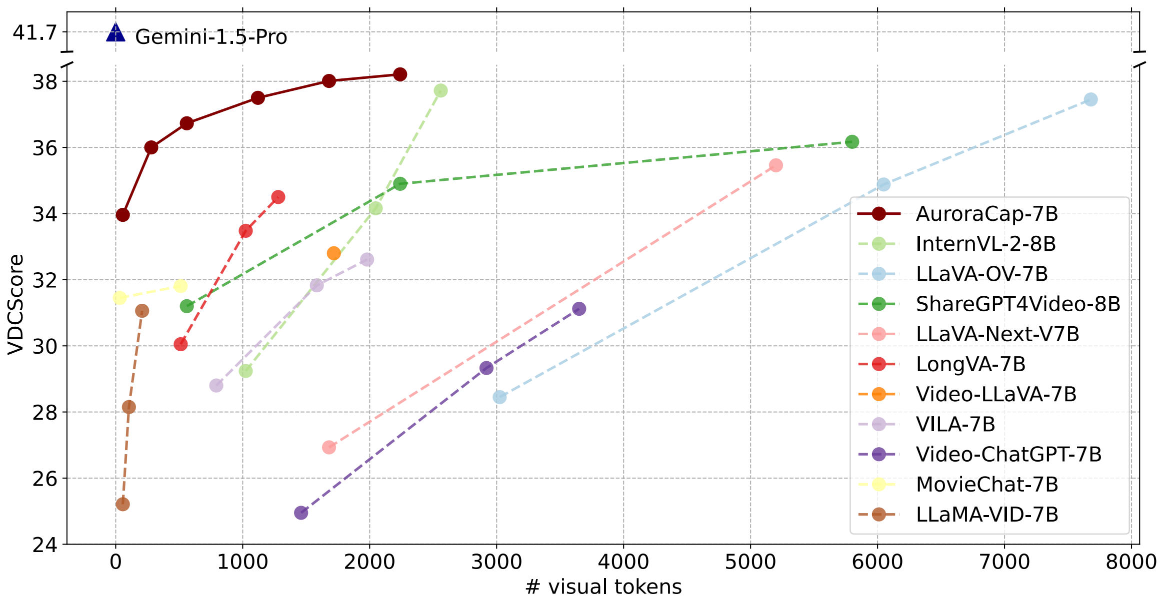


Figure 3.4: Comparison between various models with different number of visual tokens input on VDC. For Gemini-1.5-Pro, we only report the performance. We manage the number of visual tokens by managing token merging for AuroraCap, and manage the number of frames for others. AuroraCap achieves a much better VDCScore than all other models given a certain compression in the number of visual tokens kept and indeed approaches the performance of Gemini-1.5-Pro.

3.2.2 Benchmarking Video Detailed Captioning

To our knowledge, no standard evaluation benchmarks have been established for detailed video captioning. To advance this field, we assess several baselines on our proposed VDC. As illustrated in Table 3.2, we present a quantitative comparison between our AuroraCap with existing state-of-the-art LMMs. We compare the VDCScore with both rule-based and model-based caption metrics with AuroraCap performing well. BLEU [109], CIDEr [140], ROUGE-L [86] and METEOR [6] are included as representative rule-based metrics. For model-based metrics, we also consider Video Detailed Description [75] (VDD), which employs `ChatGPT` as an evaluation assistant to compare full captions.

Table 3.3 presents VDCScore performance across various sections of structured captions within VDC. Following [21], we also incorporate vision-blind baselines. Furthermore, Figure 3.4 illustrates a schematic diagram of the performance and efficiency of different video training models. Since the comparison of model inference time under different architectures, models, deployment frameworks, and output lengths is unfair, so we used the number of visual tokens as a representation of efficiency. AuroraCap achieves superior performance in video detailed captioning while utilizing significantly fewer visual tokens than other models, fully highlighting the efficiency of AuroraCap. We also show additional experimental results with VDD metric as shown in Table 3.4. We also perform a human study using Elo ranking to supplement our evaluation and provide a more intuitive assessment of AuroraCap's performance. As depicted in Figure 3.5, VDCscore shows the better correlation with human evaluation results than VDD and ROUGE metric.

Ablation study on token merging ratio. As a core strategy of AuroraCap, token merging plays a significant role in reducing the number of visual tokens. We further study how the video detailed captioning capability is influenced by token merge ratio. We define the performance percentage as the proportion between the highest and lowest values on the entire performance curve. As shown in Figure 3.6, most models maintain satisfactory performance ($> 80\%$) even with only 0.2 of visual token kept ratio. Since AuroraCap focuses on spatial visual token merging, the temporal features introduce additional complexity to explore the token merging laws, resulting in the optimal performance may occurs at a middle level of visual token kept ratio.

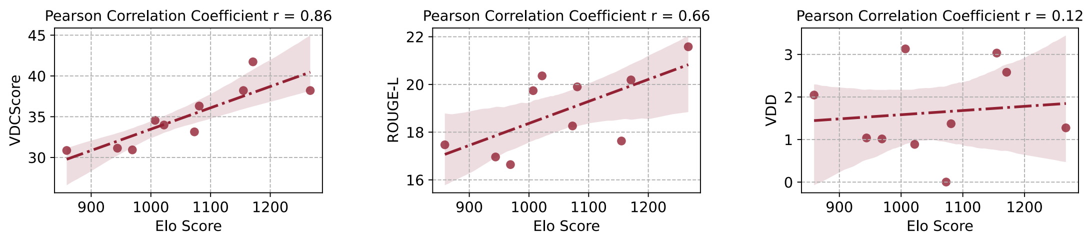


Figure 3.5: Pearson correlation analysis among three evaluation metrics— VDCSCORE, ROUGE-L [86], VDD [91], and human Elo rankings for video models. VDD reflects the score for detailed captions. VDCSCORE demonstrates the highest consistency with expert judgments, thereby reinforcing the reliability. The detailed settings are provided in the Appendix L.

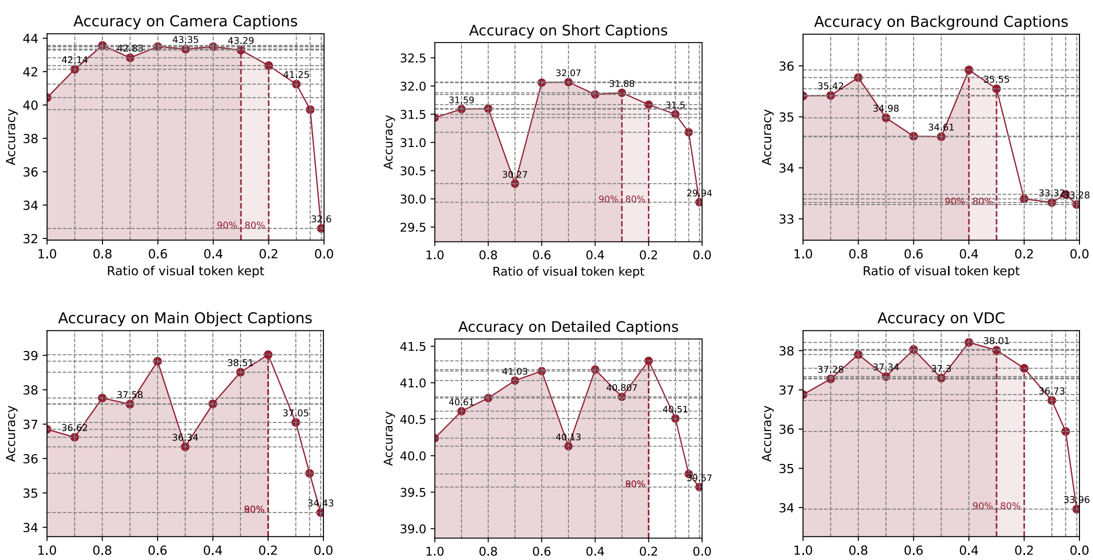


Figure 3.6: Ablation study of token merging on VDC. We found that token merging significantly reduces the number of tokens while maintaining minimal performance drop, and even showing improvement in some aspects. We highlight the token merging ratio when achieving 90% and 80% performance with the dash line and filled area.

Table 3.2: Comparison of AuroraCap with LLM-based baseline methods on VDCscore and other evaluation metrics under zero-shot setting. For each evaluation metric, we report the average value of the five structured captions in VDC. Note that VDD, CIDEr, and BELU are only the average of background and main object caption, since the values of the others are closed to zero.

Model	VDCscore		VDD		C	B@1	B@4	M	R	Elo
	Acc	Score	Acc	Score						
Gemini-1.5 Pro [117]	41.73	2.15	49.68	3.07	5.97	29.72	2.63	21.21	20.19	1,171
LLaMA-VID [82]	30.86	1.62	4.63	1.63	1.48	17.74	1.46	8.07	17.47	859
Video-ChatGPT-7B [97]	31.12	1.62	8.57	1.84	2.92	17.31	2.19	11.57	16.96	944
MovieChat-7B [127]	31.92	1.64	10.24	1.86	5.14	14.33	3.17	13.60	14.98	890
VILA-7B [87]	32.61	1.70	16.27	2.02	8.20	19.13	2.11	5.62	16.63	1,073
Video-LLaVA-7B [85]	32.80	1.72	14.14	2.00	4.43	17.20	2.32	10.36	17.53	1,007
LLaVA-1.5-7B [90]	33.98	1.76	26.71	2.33	6.63	29.80	2.54	22.79	20.36	825
LongVA-7B [171]	34.50	1.79	32.65	2.69	4.83	18.75	2.16	13.43	14.84	969
LLaVA-1.5-13B [90]	34.78	1.80	28.26	2.36	3.90	20.43	2.02	26.37	17.87	943
LLaVA-NeXT-V7B [175]	35.46	1.85	25.62	2.34	2.66	20.18	2.33	**28.17**	17.51	1,022
LLaVA-1.6-7B [91]	35.70	1.85	40.16	2.69	3.09	17.36	1.59	24.23	17.08	846
LLaVA-1.6-13B [91]	35.85	1.85	34.55	2.51	5.55	29.23	2.50	20.26	19.96	728
ShareGPT4Video-8B [23]	36.17	1.85	36.44	1.85	1.02	12.61	0.79	8.33	16.31	1,102
LLaVA-OV-7B [70]	37.45	1.94	41.83	2.70	4.09	28.34	2.84	23.98	19.59	1,155
InternVL-2-8B [27]	37.72	1.96	**48.99**	**3.03**	5.59	15.75	2.48	10.76	17.63	1,081
AuroraCap-7B	**38.21**	**1.98**	48.33	2.90	**9.51**	**30.90**	**4.06**	19.09	**21.58**	**1,267**

Table 3.3: Comparison of AuroraCap with LLM-based baseline methods on VDCscore under zero-shot structured captions setting. We consider the VDCscore of detailed captions, short captions, background captions, main object captions and camera captions. We also test the vision-blind case suggesting by [21, 137].

Model	Camera	Short	Background	Main Object	Detailed
	Acc / Score	Acc / Score	Acc / Score	Acc / Score	Acc / Score
Vicuna-v1.5-7B [29]	21.68 / 1.12	23.06 / 1.17	22.02 / 1.15	22.64 / 1.16	23.09 / 1.20
Llama-3.1-8B [39]	17.83 / 1.00	17.90 / 1.02	19.52 / 1.10	19.57 / 1.10	20.10 / 1.22
Gemini-1.5 Pro [117]	38.68 / 2.05	35.71 / 1.85	43.84 / 2.23	47.32 / 2.41	43.11 / 2.22
LLaMA-VID [82]	39.47 / 2.10	29.92 / 1.56	28.01 / 1.45	31.24 / 1.59	25.67 / 1.38
Video-ChatGPT-7B [97]	37.46 / 2.00	29.36 / 1.56	33.68 / 1.70	30.47 / 1.60	24.61 / 1.26
MovieChat-7B [127]	37.25 / 1.98	32.55 / 1.59	28.99 / 1.54	31.97 / 1.64	28.82 / 1.46
VILA-7B [87]	34.33 / 1.83	30.40 / 1.55	35.15 / 1.80	33.38 / 1.72	29.78 / 1.58
Video-LLaVA-7B [85]	37.48 / 1.97	30.67 / 1.63	32.50 / 1.70	36.01 / 1.85	27.36 / 1.43
LLaVA-1.5-7B [90]	38.38 / 2.04	28.61 / 1.51	34.86 / 1.79	34.62 / 1.76	33.43 / 1.73
LongVA-7B [171]	35.32 / 1.90	31.94 / 1.63	36.39 / 1.85	40.95 / 2.11	27.91 / 1.48
LLaVA-1.5-13B [90]	38.97 / 2.07	30.89 / 1.60	34.79 / 1.78	36.27 / 1.84	33.00 / 1.74
LLaVA-NeXT-V7B [175]	39.73 / 2.10	30.63 / 1.60	36.54 / 1.88	36.54 / 1.88	33.84 / 1.77
LLaVA-1.6-7B [91]	36.50 / 1.93	31.91 / 1.65	37.58 / 1.92	36.03 / 1.85	36.47 / 1.89
LLaVA-1.6-13B [91]	35.61 / 1.86	31.90 / 1.66	**38.90** / **1.99**	36.65 / 1.87	36.18 / 1.89
ShareGPT4Video-8B [23]	33.28 / 1.76	**39.08** / **1.94**	35.77 / 1.81	37.12 / 1.89	35.62 / 1.84
LLaVA-OV-7B [70]	37.82 / 2.02	32.58 / 1.70	37.43 / 1.92	38.21 / 1.96	41.20 / 2.13
InternVL-2-8B [27]	39.08 / 2.11	33.02 / 1.74	37.47 / 1.89	**44.16** / **2.22**	34.89 / 1.82
AuroraCap-7B	**43.50** / **2.27**	32.07 / 1.68	35.92 / 1.84	39.02 / 1.97	**41.30** / **2.15**

Table 3.4: Comparison of AuroraCap with LLM-based baseline methods on VDD [91] under zero-shot structured captions setting. We consider the VDD [91] of detailed captions, short captions, background captions, main object captions and camera captions.

Model	Camera	Short	Background	Main Object	Detailed
	Acc / Score	Acc / Score	Acc / Score	Acc / Score	Acc / Score
Gemini-1.5 Pro [117]	18.89 / 2.115	16.91 / 1.572	57.73 / 3.263	41.64 / 2.886	2.581 / 0.330
LLaMA-VID [82]	28.28 / 2.513	1.034 / 1.042	6.198 / 1.895	3.063 / 1.366	2.046 / 0.304
Video-ChatGPT-7B [97]	16.00 / 2.175	4.173 / 1.032	14.14 / 2.273	3.001 / 1.423	1.038 / 0.192
VILA-7B [87]	4.005 / 1.751	2.087 / 0.233	22.45 / 2.385	10.10 / 1.672	1.015 / 0.262
Video-LLaVA-7B [85]	20.00 / 2.336	3.193 / 1.064	17.17 / 2.253	11.11 / 1.765	3.130 / 0.316
LLaVA-1.5-7B [90]	26.88 / 2.515	1.222 / 0.793	36.50 / 2.725	16.92 / 1.937	0.694 / 0.276
LongVA-7B [171]	17.00 / 2.204	1.016 / 0.794	50.00 / 3.203	15.31 / 2.196	0.002 / 0.247
LLaVA-1.5-13B [90]	32.65 / 2.662	2.836 / 0.922	37.55 / 2.749	18.98 / 1.978	0.688 / 0.275
LLaVA-NeXT-V7B [175]	29.81 / 2.645	1.913 / 0.957	33.23 / 2.692	18.02 / 1.999	0.887 / 0.279
LLaVA-1.6-7B [91]	21.11 / 2.229	8.696 / 1.146	53.31 / 3.105	27.01 / 2.286	1.282 / 0.279
LLaVA-1.6-13B [91]	21.56 / 2.199	9.798 / 1.206	39.37 / 2.692	29.73 / 2.329	1.287 / 0.271
ShareGPT4Video-8B [23]	33.28 / 1.768	4.908 / 0.986	35.77 / 1.813	37.12 / 1.899	**3.213** / **0.752**
LLaVA-OV-7B [70]	17.11 / 2.086	**11.15** / **1.277**	55.82 / 3.149	27.84 / 2.258	1.372 / 0.249
InternVL-2-8B [27]	29.00 / 2.545	4.041 / 1.079	**63.64** / **3.446**	34.34 / **2.627**	3.032 / 0.394
AuroraCap-7B	**49.40** / **3.141**	3.313 / 0.886	59.52 / 3.261	**37.14** / 2.533	1.275 / 0.295

Chapter 4
CONCLUSION

In this paper, we first introduce AuroraCap, a efficient video detailed captioner based on large multimodal model. By leveraging the token merging strategy, we significantly reduce the computational overhead without compromising performance. We also present VDC, a novel video detailed captioning benchmark designed to evaluate comprehensive and coherent textual descriptions of video content. For better evaluating, We propose VDCscore , a new LLM-assisted metric with divide-and-conquer strategy. Our extensive evaluation on various video and image captioning benchmarks demonstrated that AuroraCap achieves competitive results, even outperforming state-of-the-art models in some tasks. We also conduct thorough ablation studies to validate the effectiveness of token merging and other aspects of our model. We found that the current model performs poorly in terms of the trade-off between performance and the scale of input tokens. Additionally, there is still room for improvement in camera handling and detailed captioning. We hope that VDC can bring new insights to the video detailed captioning task.

BIBLIOGRAPHY

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 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*, pages 382–398. Springer, 2016.

[4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pages 1728–1738, 2021.

[6] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.

[7] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024.

[8] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023.

[9] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[10] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019.

[11] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your ViT but faster. In *The Eleventh International Conference on Learning Representations*, 2022.

[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[13] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024.

[14] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.

[15] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

[16] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[17] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

[18] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024.

[19] Liang Chen, Shuai Bai, Wenhao Chai, Weichu Xie, Haozhe Zhao, Leon Vinci, Junyang Lin, and Baobao Chang. Multimodal representation alignment for image generation: Text-image interleaved control is easier than you think. *arXiv preprint arXiv:2502.20172*, 2025.

[20] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. *arXiv preprint arXiv:2403.06764*, 2024.

[21] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[22] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[23] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

[24] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024.

[25] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[26] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. Chinesefoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.

[27] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[28] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[29] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An

open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[30] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023.

[31] Dawei Dai, YuTang Li, YingGe Liu, Mingming Jia, Zhang YuanHui, and Guoyin Wang. 15m multimodal facial image-text dataset, 2024.

[32] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[34] Jie Deng, Wenhao Chai, Jianshu Guo, Qixuan Huang, Wenhao Hu, Jenq-Neng Hwang, and Gaoang Wang. Citygen: Infinite and controllable 3d city layout generation. *arXiv preprint arXiv:2312.01508*, 2023.

[35] Jie Deng, Wenhao Chai, Junsheng Huang, Zhonghan Zhao, Qixuan Huang, Mingyan Gao, Jianshu Guo, Shengyu Hao, Wenhao Hu, Jenq-Neng Hwang, et al. Citycraft: A real crafter for 3d city generation. *arXiv preprint arXiv:2406.04983*, 2024.

[36] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[37] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Zhengfang Duanmu, Wentao Liu, Zhongling Wang, and Zhou Wang. Quantifying visual image quality: A bayesian view. *Annual Review of Vision Science*, 7:437–464, 2021.

[39] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[40] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*, 2016.

[41] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[42] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

[43] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015.

[44] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. Imageinwords: Unlocking hyper-detailed image descriptions. *arXiv preprint arXiv:2405.02793*, 2024.

[45] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[46] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[47] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[48] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

[49] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[50] Mingfei Han, Xiaojun Chang, Heng Wang, and Linjie Yang. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*, 2023.

[51] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[52] Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*, 2023.

[53] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016.

[54] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[55] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

[56] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020.

[57] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[58] Jitesh Jain, Jianwei Yang, and Humphrey Shi. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*, 2023.

[59] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

[60] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[61] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024.

[62] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024.

[63] Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254, 2021.

[64] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[65] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.

[66] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[67] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[68] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[69] Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120, 2022.

[70] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[71] Boyi Li, Ligeng Zhu, Ran Tian, Shuhan Tan, Yuxiao Chen, Yao Lu, Yin Cui, Sushant Veer, Max Ehrlich, Jonah Philion, et al. Wolf: Captioning everything with a world summarization framework. *arXiv preprint arXiv:2407.18908*, 2024.

[72] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[73] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[74] Junyan Li, Delin Chen, Tianle Cai, Peihao Chen, Yining Hong, Zhenfang Chen, Yikang Shen, and Chuang Gan. Flexattention for efficient high-resolution vision-language models. *arXiv preprint arXiv:2407.20228*, 2024.

[75] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[76] Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023.

[77] Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*, 2023.

[78] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024.

[79] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. Adding chinese captions to images. In *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pages 271–275, 2016.

[80] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.

[81] Xirui Li, Chao Ma, Xiaokang Yang, and Ming-Hsuan Yang. Vidtome: Video token merging for zero-shot video editing. *arXiv preprint arXiv:2312.10656*, 2023.

[82] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023.

[83] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023.

[84] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[85] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[86] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[87] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533*, 2023.

[88] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[89] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8611–8619, 2018.

[90] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[91] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[92] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[93] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[94] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[95] Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. Vista-llama: Reliable video narrator via equal distance to visual tokens. *arXiv preprint arXiv:2312.08870*, 2023.

[96] Yunchuan Ma, Laiyun Qing, Guorong Li, Yuankai Qi, Quan Z Sheng, and Qingming Huang. Retrieval enhanced zero-shot video captioning. *arXiv preprint arXiv:2405.07046*, 2024.

[97] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[98] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[99] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[100] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[101] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[102] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[103] Mixkit. Free stock video clips, motion graphics, & music tracks, 2024. Accessed: 2024-08-24.

[104] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021.

[105] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer, 2022.

[106] Nous Research. Nous-hermes-2-yi-34b, 2023. Accessed: 2024-08-29.

[107] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

[108] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[109] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[110] Pexels. Free imageds and video clips, 2024. Accessed: 2024-08-24.

[111] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13018–13028, 2021.

[112] Pixabay. Stunning free images & royalty-free stock, 2024. Accessed: 2024-08-24.

[113] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[114] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[115] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[116] Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

[117] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[118] Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. Testa: Temporal-spatial token aggregation for long-form video-language understanding. *arXiv preprint arXiv:2310.19060*, 2023.

[119] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 433–440, 2013.

[120] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[121] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[122] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[123] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024.

[124] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[125] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[126] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[127] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023.

[128] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024.

[129] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, 2017.

[130] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. *arXiv preprint arXiv:2404.04346*, 2024.

[131] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13878–13888, 2021.

[132] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[133] Qwen team. Qwen2-vl. 2024.

[134] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022.

[135] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[136] Yunjie Tian, Tianren Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. Chatterbox: Multi-round multimodal referring and grounding. *arXiv preprint arXiv:2401.13307*, 2024.

[137] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

[138] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

[139] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015.

[140] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[141] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

[142] Han Wang, Yanjie Wang, Yongjie Ye, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. *arXiv preprint arXiv:2403.16558*, 2024.

[143] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[144] Jie Wang, Tao Ji, Yuanbin Wu, Hang Yan, Tao Gui, Qi Zhang, Xuanjing Huang, and Xiaoling Wang. Length generalization of causal transformers without position encoding. *arXiv preprint arXiv:2404.12224*, 2024.

[145] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.

[146] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019.

[147] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

[148] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023.

[149] Wenhao Wu. Freeva: Offline mllm as training-free video assistant. *arXiv e-prints*, pages arXiv–2405, 2024.

[150] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

[151] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

[152] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[153] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*, pages 38728–38748. PMLR, 2023.

[154] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[155] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*, 2024.

[156] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024.

[157] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022.

[158] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022.

[159] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[160] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021.

[161] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022.

[162] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 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*, pages 10714–10726, 2023.

[163] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[164] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, 2023.

[165] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[166] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[167] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[168] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[169] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[170] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[171] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[172] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[173] Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*, 2024.

[174] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[175] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.

[176] Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, et al. Distilling vision-language models on millions of videos. *arXiv preprint arXiv:2401.06129*, 2024.

[177] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Li Boyi, Shengyu Hao, Shidong Cao, Tian Ye, Jenq-Neng Hwang, and Gaoang Wang. See and think: Embodied agent in virtual environment. *arXiv preprint arXiv:2311.15209*, 2023.

[178] Zhonghan Zhao, Wenhao Chai, Xuan Wang, Ke Ma, Kewei Chen, Dongxu Guo, Tian Ye, Yanting Zhang, Hongwei Wang, and Gaoang Wang. Steve series: Step-by-step construction of agent systems in minecraft. *Computer Vision and Pattern Recognition Workshop*, 2024.

[179] Zhonghan Zhao, Kewei Chen, Dongxu Guo, Wenhao Chai, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*, 2024.

[180] Zhonghan Zhao, Ke Ma, Wenhao Chai, Xuan Wang, Kewei Chen, Dongxu Guo, Yanting Zhang, Hongwei Wang, and Gaoang Wang. Do we really need a complex agent system? distill embodied agent into a single model. *arXiv preprint arXiv:2404.04619*, 2024.

[181] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*, 2021.

[182] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[183] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. *arXiv preprint arXiv:2404.01297*, 2024.

[184] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2023.

Appendix A
RELATED WORKS

Video Captioning The general goal of video captioning is understanding a video and describing it with natural language. Unlike image captioning, video captioning requires the description also on the temporal dimension (*e.g.,* human action, camera and object movement). The current datasets including videos in the domain of human activities [12], cooking scene [182, 56], movie [127], and open domain [17]. In Table 1.1, we show the summary of current video captioning benchmarks. Most of them are with short caption, which is not suitable for evaluating video detailed captioning task. There are several widely used metrics to evaluate the correctness of generated caption, such as BLEU [109], ROUGE-L [86], CIDEr [140], SPICE [3], and BERTScore [174]. Dense Video Captioning [162, 183] is also a captioning task but further requires localizing the events temporally. In this paper, we focus on the detailed description of a short video clip, where there are no scene changes or camera switches.

Large Multimodal Models for Video With the develop of LLMs and LMMs [97, 172, 82, 127, 169, 128, 176, 183, 133, 59, 9, 123, 13, 28, 74, 15, 19], many recent works have explored adapting them into video understanding field (*e.g.,* Video-LLaMA [82], Video-LLaVA [85], VideoChat [75], Vista-LLaMA [95], LLaVA-Hound [173], Koala [130], Elysium [142], and MovieChat [127, 128]). Thanks to this flexible design, the models can combine pretrained knowledge with minimal trainable parameters. Instead of requiring large-scale training, using only a small amount of high-quality training data can even achieve better results. Most of the existing models use additional parameters for temporal modeling. There are also some interesting observations that we can actually build advancing LMMs without additional parameters for temporal modeling [155, 87, 91, 170, 118] or even without further training with video-text data [149]. Recent works (*e.g.,* FreeVA [149], LLaVA-Next [91], VLIA [87], and PLLaVA [155]) also find that the vanilla LLaVA-like model pretrained on high-quality image instruction data can also be a strong video understanding model. FreeVA further observe that using existing video instruction tuning data like Video-ChatGPT 100K [97] to tune LMMs may not necessarily lead to improvements. As the concurrent work, we also observe this phenomenon and proceeded to develop a video detailed captioning baseline training based on the LLaVA architecture.

As for the benchmark , inspired by LLaVA [92], Video-ChatGPT [97] introduces a 100K video clips with text instructions with the first vLMMs benchmark evaluation system powered by LLMs. MovieChat-1K [127] and CinePile [116] are question-answering based benchmark for long-form video understanding. Shot2Story20K [50] comprises videos with 2 to 8 shots each sourced for our dataset from the public video benchmark HDvila100M [157]. However, currently there is no video benchmark available to evaluate video detailed captioning tasks like IIW [44] did in image captioning field. In this paper, our work fills this gap.

Appendix B

BENCHMARK COMPARISON

We compare our proposed VDCwith some examples from several video captioning benchmarks as shown in Figure B.1, Figure B.2, and Figure B.3. The corresponding captions are shown as followings:

Figure B.1: Video example of MSR-VTT [154] benchmark.

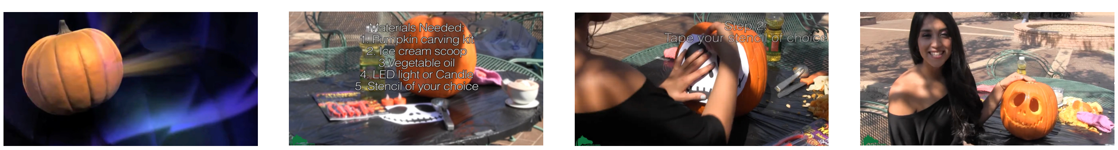


Figure B.2: Video example of VATEX [146] benchmark.

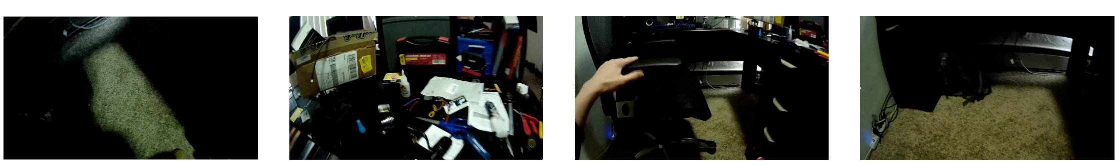


Figure B.3: Video example of VDC benchmark sampled from Ego4D [47].

Benchmark Caption (for VDC we show only the detailed caption)

MSR-VTT B.1 (4 words) Teams are playing soccer.

VATEX B.2 (13 words) A woman instructs and demonstrates how to remove the insides of a pumpkin.

VDC B.3 (618 words) The video opens with an intimate close-up of a surface adorned with vibrant green moss and intricate lichen, initially evoking the serene beauty of a natural landscape. This organic imagery quickly transitions, revealing that the mossy surface is actually part of a motorcycle or vehicle's engine compartment, creating a striking contrast between the lush textures of nature and the cold, hard lines of mechanical components. As the camera angle shifts, the viewer is drawn deeper into the engine compartment, where the interplay of moss overgrowth on various machinery introduces a fascinating blend of organic life and industrial elements, highlighting the unexpected coexistence of nature and technology.

The perspective then zooms in, accentuating the rich details of the mossy growth, which clings tenaciously to the metallic surfaces, while a dark cavity beneath hints at the complexity of the machinery. The reflective metallic surfaces glint in the light, further enhancing the visual contrast and inviting the viewer to explore this unique juxtaposition. Suddenly, the scene shifts dramatically, with rapid camera motion creating a vibrant blur of colors and shapes, transforming the previously detailed views into a chaotic whirlwind, suggesting a swift movement through the intricate landscape of the engine compartment. As the motion blur begins to dissipate, the viewer is presented with a clearer image of a light-colored, textured surface, where blurred mechanical components can be discerned, indicating a deceleration in movement. The camera stabilizes, revealing a rough-textured floor or ground that suggests an indoor or industrial environment, characterized by a sense of organized chaos. The scene transitions to a detailed examination of a cluttered workspace filled with tangled wires, casings, and components in a variety of colors, emphasizing the disorganized state of electronic or mechanical internals, possibly during a maintenance or repair process.

The perspective shifts once more, showcasing darker, textured surfaces juxtaposed against lighter insulating materials, with hidden metallic elements peeking through, suggesting another angle within this same cluttered interior space. A human hand enters the frame, reaching out to interact with the components, signaling an active workspace filled with purpose. As the scene expands, additional hands join the fray, actively manipulating various objects within the crowded environment, signifying an ongoing task or collaborative effort amidst the complex array of components and materials.

The atmosphere is imbued with a sense of urgency and engagement, as the camera captures the dynamic interactions of the individuals working together. The camera work remains fluid and dynamic, featuring a mix of close-up shots that highlight the intricate details of the components and wider angles that provide context to the bustling environment. The slightly shaky nature of the shots adds a layer of realism and immersion, drawing the viewer into the heart of the action. The low light conditions create a moody ambiance, with shadows dancing across the surfaces, enhancing the visual depth and interest of the scene. Overall, the video encapsulates a vivid portrayal of the intersection between nature and machinery, as well as the collaborative spirit of those engaged in the intricate task of maintenance and repair within this unique setting.

Appendix C

TOKEN MERGING VISUALIZATION

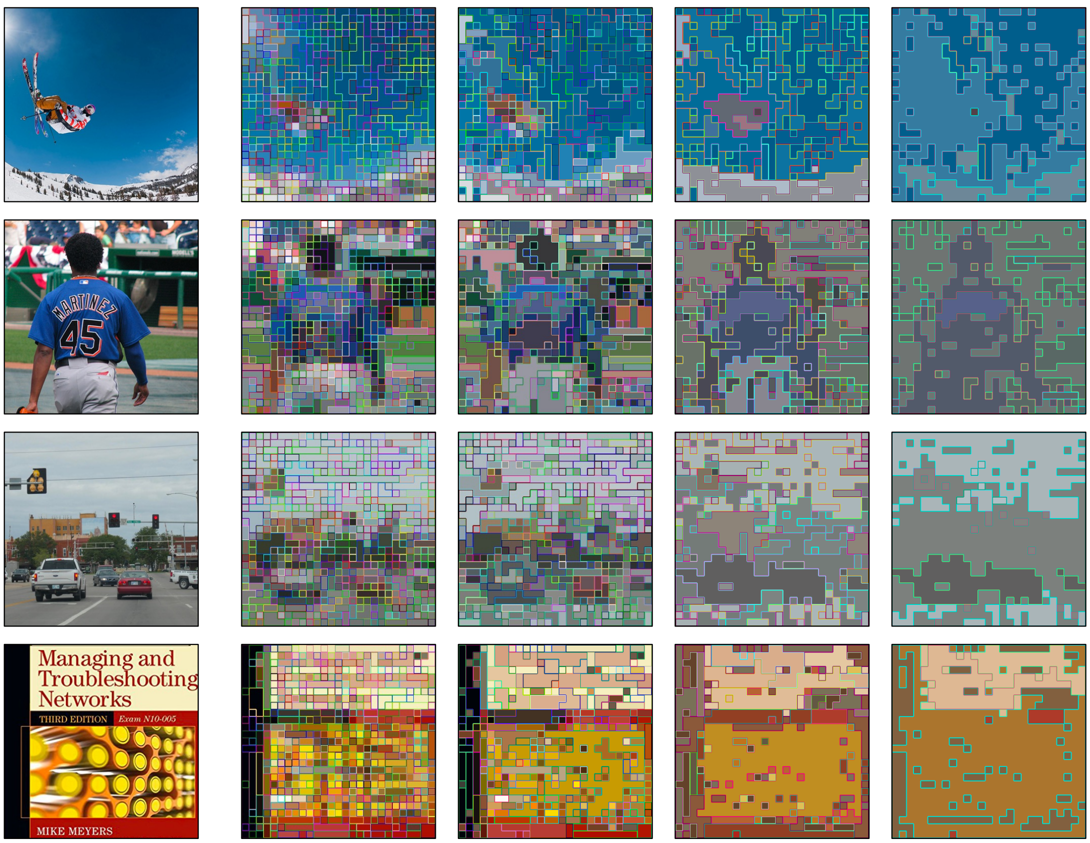


Figure C.1: Token merging visualization. From top to bottom, the image IDs are COCO:COCO-train2014-000000247906, VG:2331508, SA-1B:sa-393200, OCR-VQA:71789111. From left to right, the number of tokens representing the images are 490, 154, 18, and 6.

The number of tokens r merged in each layer of the transformer in ViT is calculated by the following formula:

$$r = (1 - \text{ratio}) \cdot (W \cdot H/P^2)/L \tag{C.1}$$

where W, H is the width and the height of the input image, P is the patch size, L is the number of layers in the transformer, ratio is the proportion of tokens to be kept. The merging is evenly distributed across all layers of the transformer.

We show token merging visualization examples from COCO [88], VG [67], SA-1B [64], and OCR-VQA [102] in Figure C.1, since we only conduct frame-wise operation. Features extracted by CLIP ViT-H model [1]. We conduct image captioning for the examples given final token numbers by 7B AuroraCap as following. We remove line breaks in generated caption for a better reading experience.

 # Token Describe this image in detail. (Figure C.1 Row 1)

490 In the heart of a winter wonderland, a skier takes flight, suspended in mid-air against the backdrop of a clear blue sky. The skier, clad in a vibrant red and white jacket, is captured in a moment of pure exhilaration. The skier's body is angled towards the left side of the image, with their skis crossed in an impressive display of skill and control. The skier's helmet, a stark contrast to the surrounding snow, is adorned with a pink and white design, adding a touch of color to the otherwise monochrome landscape. The skier's goggles, reflecting the bright sunlight, are a testament to the intensity of the sport. Beneath the skier, the snow-covered mountain slopes stretch out, dotted with trees that stand as silent spectators to the skier's daring feat. The image is a freeze-frame of action and adventure, encapsulating the thrill and beauty of winter sports.

154 In the midst of a clear blue sky, a skier is captured in a moment of thrilling action. The skier, clad in a vibrant red and white jacket, is suspended in mid-air, performing an impressive trick. The skier's body is parallel to the ground, creating a sense of motion and excitement. The skier's skis, a striking combination of red and white, are crossed in front of them, adding to the dynamic nature of the image. The skier's helmet, a necessary safety measure, is white and contrasts with the colorful jacket. The backdrop to this action-packed scene is a breathtaking snowy mountain landscape. The mountain's peak is visible in the distance, adding depth to the image. The overall composition of the image suggests a high-speed descent down the mountain, with the skier skillfully navigating the terrain. This image

[1]Huggingface Model: https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14-378

encapsulates the exhilarating sport of skiing, showcasing not only the skill of the skier but also the stunning beauty of the mountain landscape.

18 In the heart of a snowy mountain, a daring skier takes flight. Dressed in a vibrant red and white jacket, black pants, and a protective white helmet, the skier is captured mid-air, performing an impressive trick. The skier's body is parallel to the ground, arms outstretched, as if embracing the thrill of the moment. The skier's skis, a striking combination of red and white, are pointed upwards towards the clear blue sky. The skier's poles, held firmly in their hands, add to the dynamic nature of the scene. The backdrop to this exhilarating moment is a breathtaking snowy mountain landscape. The mountain's peak, blanketed in snow, stretches out into the distance. The clear blue sky above is devoid of clouds, adding to the serene beauty of the scene. This image captures not just a moment, but a story of adventure, thrill, and the sheer joy of skiing.

6 The image captures a thrilling moment of a snowboarder in mid-air, performing an impressive trick. The snowboarder, clad in a vibrant red and white jacket, black pants, and a protective helmet, is the central figure in this dynamic scene. The snowboarder is holding onto a rope with one hand, suggesting that they are being pulled up the mountain by a snowmobile, a common practice in snowboarding to gain speed and momentum. The rope extends from the top left corner of the image, creating a sense of movement and direction. The backdrop of the image is a breathtaking snowy mountain landscape, complete with trees and other natural elements. The snowboarder is positioned in the center of the image, drawing the viewer's attention to their daring stunt. The image is a testament to the exhilarating sport of snowboarding and the skill and courage of the athletes who participate in it. It's a freeze-frame of an adrenaline-fueled moment, encapsulating the thrill and excitement of snowboarding.

 # Token What is the number in the picture? (Figure C.1 Row 2)

490 The number in the picture is "45". It is displayed on the back of the jersey that the person is wearing.

154 The number in the picture is "45".

18 The number in the picture is "45". It is displayed on the back of the baseball jersey worn by the player in the image.

6 The number in the picture is "45". It is displayed in a large, bold font on the jersey of the baseball player in the image.

 # Token What is the relative position of the white and red cars in the picture?

Which one in the left? (Figure C.1 Row 3)

490 The white car is positioned to the left of the red car in the image. The white car is closer to the foreground and is stopped at the traffic light, while the red car is further back, also stopped at the same traffic light. The relative position of the cars is such that the white car is in the left lane and the red car is in the right lane, both facing the same direction of traffic.

154 The white car is positioned to the left of the red car in the image. The red car is further back on the road, indicating it is behind the white car. This is determined by the perspective of the image, where the white car is closer to the foreground and the red car is further away, both in the same lane and moving in the same direction.

18 The white car is on the left side of the image, and the red car is on the right side. The white car is closer to the foreground, while the red car is further away, positioned behind the white car.

6 The white car is on the left side of the image, while the red car is on the right side. The white car is positioned closer to the foreground, while the red car is further back, near the intersection.

Appendix D
ABLATION STUDIES

In this section, we conduct the ablation studies in several aspects including token merging, slowfast inference, efficiency, training strategy, and backbone selection.

Token merging. As a core strategy of AuroraCap, token merging plays a significant role in reducing the number of visual tokens. We conduct extensive ablation studies to explore the impact of the token kept ratio R_{vtk} in terms of performance across multiple tasks including image captioning, visual question answering, video captioning, and video question answering as shown in Figure D.3 and Figure D.4. We define the performance percentage as the proportion between the highest and lowest values on the entire performance curve. We identify the minimum retention thresholds for achieving 90% and 80% performance. As shown in Figure D.3, while the performance of AuroraCap generally declines with fewer visual tokens across most benchmarks, it remains relatively stable at higher retention levels. Most models maintain satisfactory performance ($> 80\%$) even with only 0.4 of R_{vtk}, highlighting the efficiency of our approach. Visual token retention thresholds vary by task complexity, with more visually demanding tasks needing higher retention of visual tokens. For instance, CIDEr [140] on COCO-Cap [141] maintains over 90% performance with an R_{vtk} of 0.3, whereas accuracy on GQA [54] drops to 90% when the R_{vtk} is reduced to 0.8. Unlike image understanding, the optimal performance across most video understanding benchmarks occurs at a relatively low R_{vtk} as depicted in Figure D.4. And for MSR-VTT [154], VATEX [146], and ActivityNet-QA [166], even achieve better results at extremely low R_{vtk} (< 0.1). It indicates that comparing to image, video input have higher redundancy. Note that AuroraCap focuses on spatial visual token merging, while the temporal features introduce additional complexity to explore the token merging laws. Appendix C shows more calculation details and the visualization results of token merging.

Slowfast inference. Inspired by Slowfast-LLaVA [156], we explore whether combining frames with low and high R_{vtk} can enhance performance. In practice, we don't conduct token merging in the first frame and concatenate them with the merged tokens from subsequent frames. We apply this strategy to both video captioning and video question answering tasks, comparing performance with and without the inclusion of full first-frame visual tokens. As illustrated in Table D.1, slowfast inference brings marginal performance improvement in video question answering tasks or even drop in video captioning tasks but with more computing cost. Therefore, by default, we don't using slowfast inference for video detailed captioning.

We also present the performance curve with and without the inclusion of full first-frame visual tokens as the visual token kept ratio varies during inference across multiple video understanding tasks. As illustrated in Figure D.5 and Figure D.6, despite the inclusion of full first-frame visual tokens, slowfast inference does not consistently result in significantly positive effects on performance. In some cases, the incorporation of full first-frame visual tokens even worsens performance degradation as the kept ratio decreases, particularly in video captioning tasks.

Table D.1: Ablation on slowfast inference for AuroraCap-7B. We present the average performance among different token merging ratio on various video understanding benchmarks. We show that slowfast inference brings marginal performance improvement in video question answering tasks or even drop in video captioning tasks but with more computing cost.

Setting	MSR-VTT					VATEX					ANet	MSVD	MSRVTT
	$\overline{\text{C}}$	$\overline{\text{B@1}}$	$\overline{\text{B@4}}$	$\overline{\text{M}}$	$\overline{\text{R}}$	$\overline{\text{C}}$	$\overline{\text{B@1}}$	$\overline{\text{B@4}}$	$\overline{\text{M}}$	$\overline{\text{R}}$	$\overline{\text{Acc}}$	$\overline{\text{Acc}}$	$\overline{\text{Acc}}$
w/o slowfast	26.72	53.01	17.58	21.25	46.78	28.03	52.28	15.22	16.95	38.30	58.55	56.45	37.26
w/ slowfast	26.18	51.68	17.00	21.20	46.16	28.07	52.16	15.12	16.95	38.27	59.66	55.65	38.22
Δ	-0.54	-1.33	-0.58	-0.05	-0.62	+0.04	-0.12	-0.10	-0.01	-0.03	+1.11	-0.80	+0.96

Efficiency. To assess the inference speed, we utilize the inference time per video question-answering pair in seconds (TPV) as an evaluative metric. SGLang is an accelerated serving

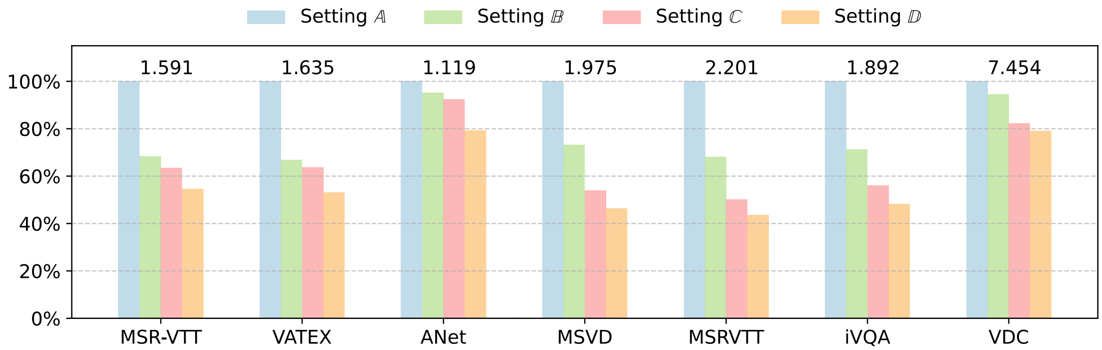


Figure D.1: Comparison between different inference settings: $\mathbb{A}$: $R_{vtk} = 1.0$, without SGLang, $\mathbb{B}$: $R_{vtk} = 0.1$, without SGLang, $\mathbb{C}$: $R_{vtk} = 1.0$, with SGLang, $\mathbb{D}$: $R_{vtk} = 0.1$, with SGLang. The number indicates the maximum inference time in seconds for each benchmark.

framework for LLMs and multimodal LLMs. We consider four settings including with or without token merging and SGLang. Figure D.1 indicates the minimum TPV achievable in each settings across seven video understanding datasets. Reducing the visual tokens and using SGLang result in excellent inference times per video question-answering pair while all the datasets with short video and question inputs. In contrast, maintaining full visual tokens or omitting the use of SGLang results in comparatively slower performance, demonstrating the superior inference efficiency of AuroraCap. For each input, we process the video at a resolution of 378×378 and sample 8 frames using single H100 GPU.

Training strategy. Alternative training strategies for the language stage of AuroraCap are less frequently explored, which is the primary focus of this section. For a fair comparison, we use the same training datasets across all settings and maintain consistent hyper-parameters. The following training settings are explored:

- **Setting $\mathbb{A}$:** End-to-end training and set R_{vtk} to 1.0.
- **Setting $\mathbb{B}$:** Following Slowfast-LLaVA [156], we retain full visual tokens in the first frame and concatenate with merged tokens using R_{vtk} of 0.1.
- **Setting $\mathbb{C}$:** End-to-end training and set R_{vtk} to 0.1.

- **Setting $\mathbb{D}$:** Most videos in the training data have no more than 8 key frames, while a subset (mainly from ShareGPT4Video [23]) contains significantly more. We first exclude this subset from end-to-end training and then use it to train solely the LLM, enhancing its ability to handle multi-frame inputs.
- **Setting $\mathbb{E}$:** Training solely the LLM match the performance of Setting $\mathbb{A}$ set R_{vtk} at 0.1.

We implement these training strategies, track training costs in H100 hours, and evaluate across various video understanding tasks. As shown in Figure D.2, while training with an R_{vtk} of 1.0 improves performance, it significantly increases training time. Surprisingly, mixing lower and higher visual token ratios during training offers no significant advantage. Training only the LLM under the two settings results in a performance drop, indicating that enhancing long video understanding still requires collaboration with the finetuning the visual encoder. Therefore, we choose Setting $\mathbb{C}$ as the final training strategy.

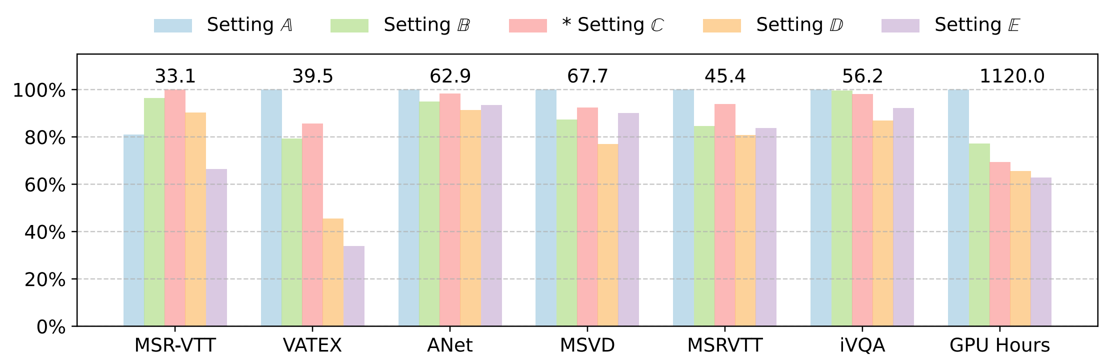


Figure D.2: Comparison between different training strategy in Language stage. We take Accuracy for Question-Answering tasks and CIDEr for captioning tasks as the evaluation metric and present the performance percentage. We choose Setting $\mathbb{C}$ as the final training strategy as shown with *. The number shows the maximum value for each benchmark.

Backbone selection. We use the the training loss among the last ten iterations in original LLaVA alignment pretraining stage to guidance the ViT and LLM backbones selection as shown in Table D.2.

Table D.2: Final training loss during pretraining stage with original LLaVA pretraining data.

ViT	ViT Size	LLM	LLM Size	Loss
facebook/dinov2-giant	1,136M	microsoft/phi-2	2.7B	3.3021
openai/clip-vit-large-patch14-336	428M	Qwen/Qwen1.5-0.5B-Chat	0.5B	3.1001
openai/clip-vit-large-patch14-336	428M	microsoft/phi-2	2.7B	2.8067
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	microsoft/phi-2	2.7B	2.7124
facebook/dinov2-giant	1,136M	lmsys/vicuna-13b-v1.5	13B	2.3895
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	internlm/internlm2-chat-7b	7B	2.3437
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	internlm/internlm2-chat-20b	20B	2.2745
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	deepseek-ai/deepseek-llm-67b-chat	67B	2.1572
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	mistralai/Mistral-7B-Instruct-v0.1	7B	2.1569
openai/clip-vit-large-patch14-336	428M	mistralai/Mixtral-8x7B-Instruct-v0.1	8x7B	2.0815
apple/DFN5B-CLIP-ViT-H-14-378	632M	lmsys/vicuna-13b-v1.5-16k	13B	2.0443
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	lmsys/vicuna-7b-v1.5-16k	7B	2.0365
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	mistralai/Mixtral-8x7B-Instruct-v0.1	8x7B	1.9889
openai/clip-vit-large-patch14-336	428M	lmsys/vicuna-7b-v1.5	7B	1.9762
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	meta-llama/Llama-2-13b-chat-hf	13B	1.9708
laion/CLIP-ViT-bigG-14-laion2B-39B-b160k	1,845M	lmsys/vicuna-13b-v1.5	13B	1.9412
apple/DFN5B-CLIP-ViT-H-14-378	632M	lmsys/vicuna-7b-v1.5-16k	7B	**1.8679**

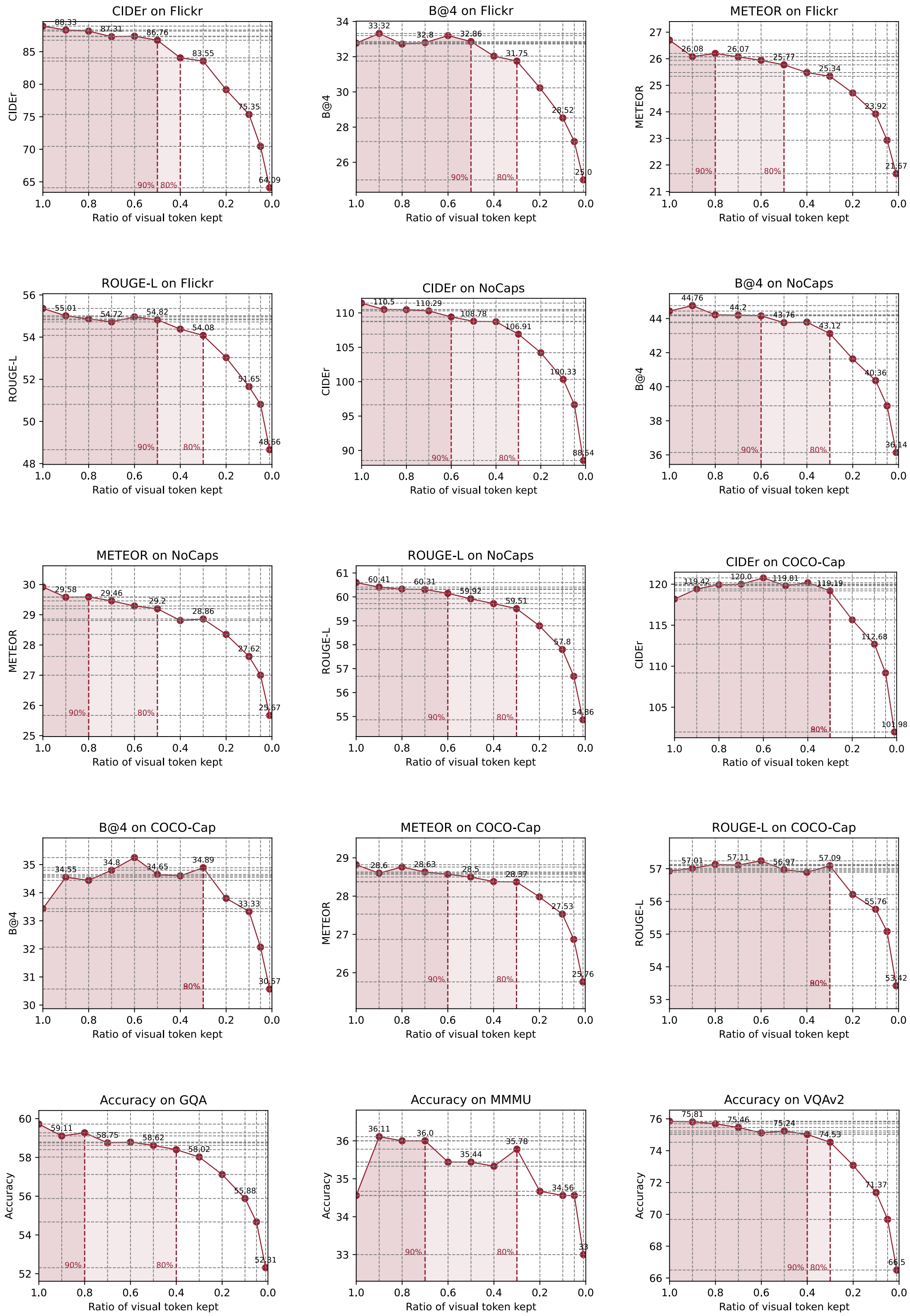


Figure D.3: Ablation study of token merging on image captioning on Flickr [165], NoCaps [1], COCO-Cap [88], visual question answering in GQA [54], MMMU [167], VQAv2 [46]. We found that token merging significantly reduces the number of tokens while maintaining

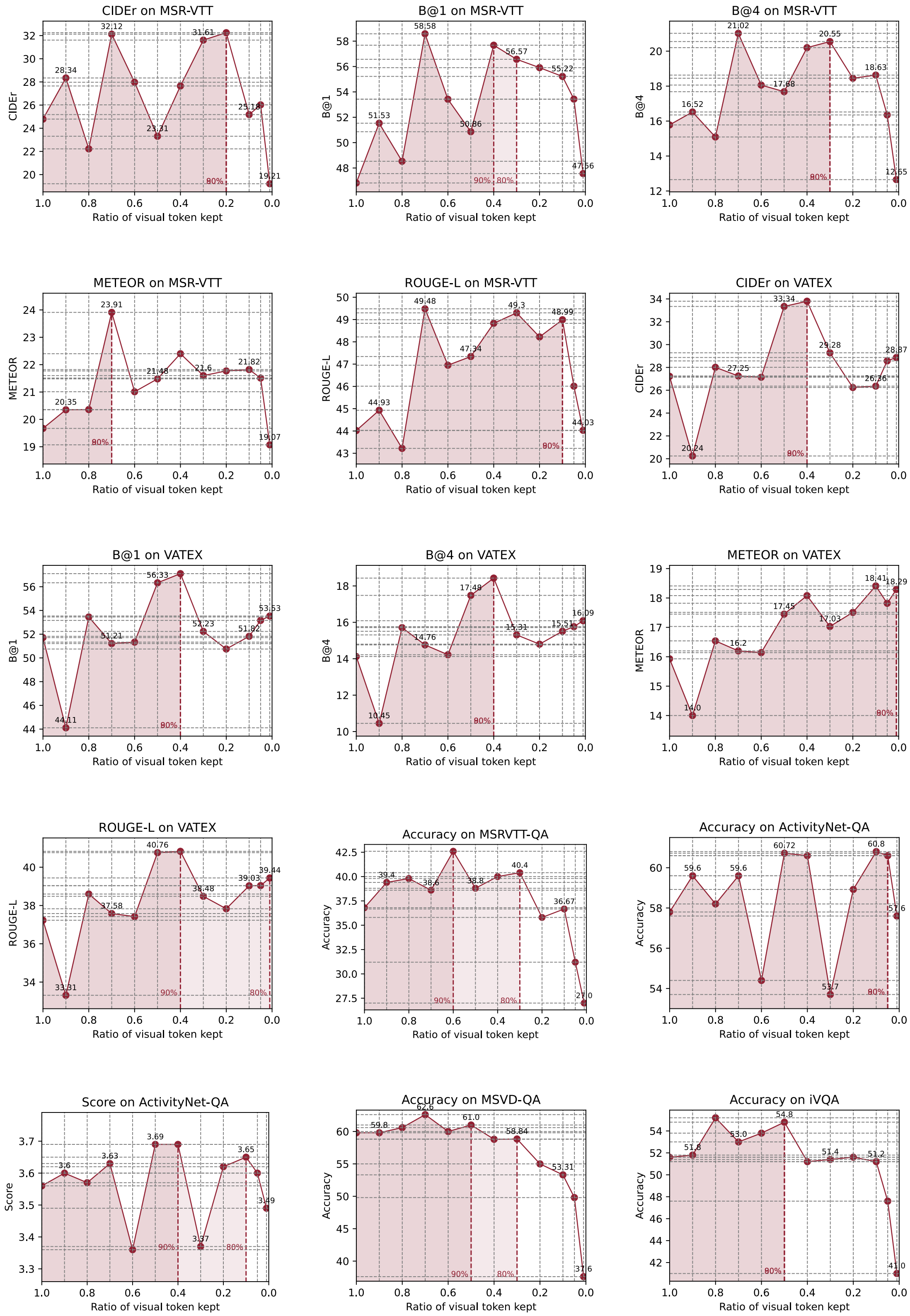


Figure D.4: Ablation study of token merging on video captioning on MSRVTT [154], VATEX [146], video question answering on MSRVTT-QA [154], ActivityNet-QA [166], MSVD-QA [152], iVQA [89]. We found that token merging significantly reduces the number

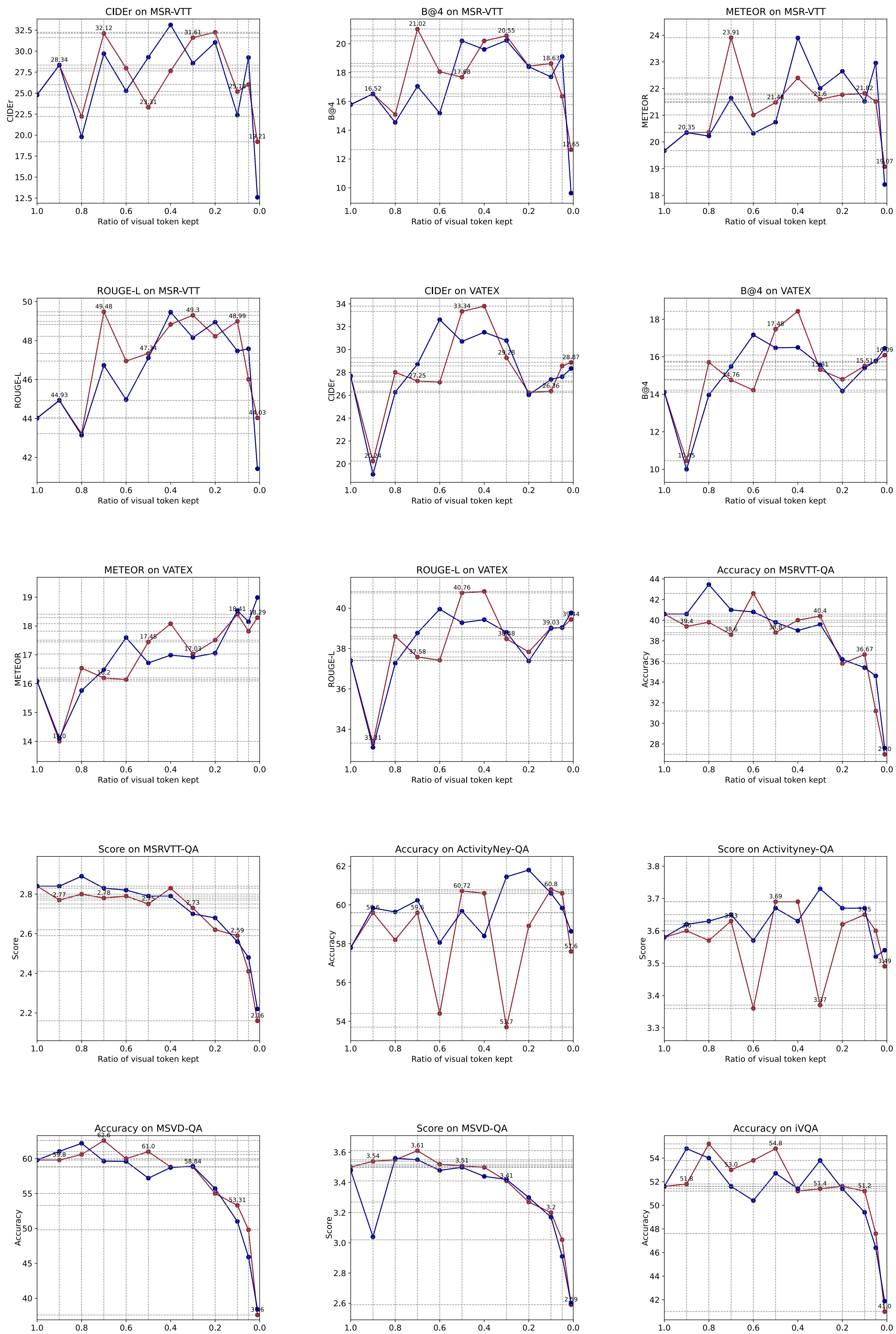


Figure D.5: Comparison between performance with and without the inclusion of full first-
frame visual tokens during inference on video captioning on MSRVTT [154], VATEX [146],

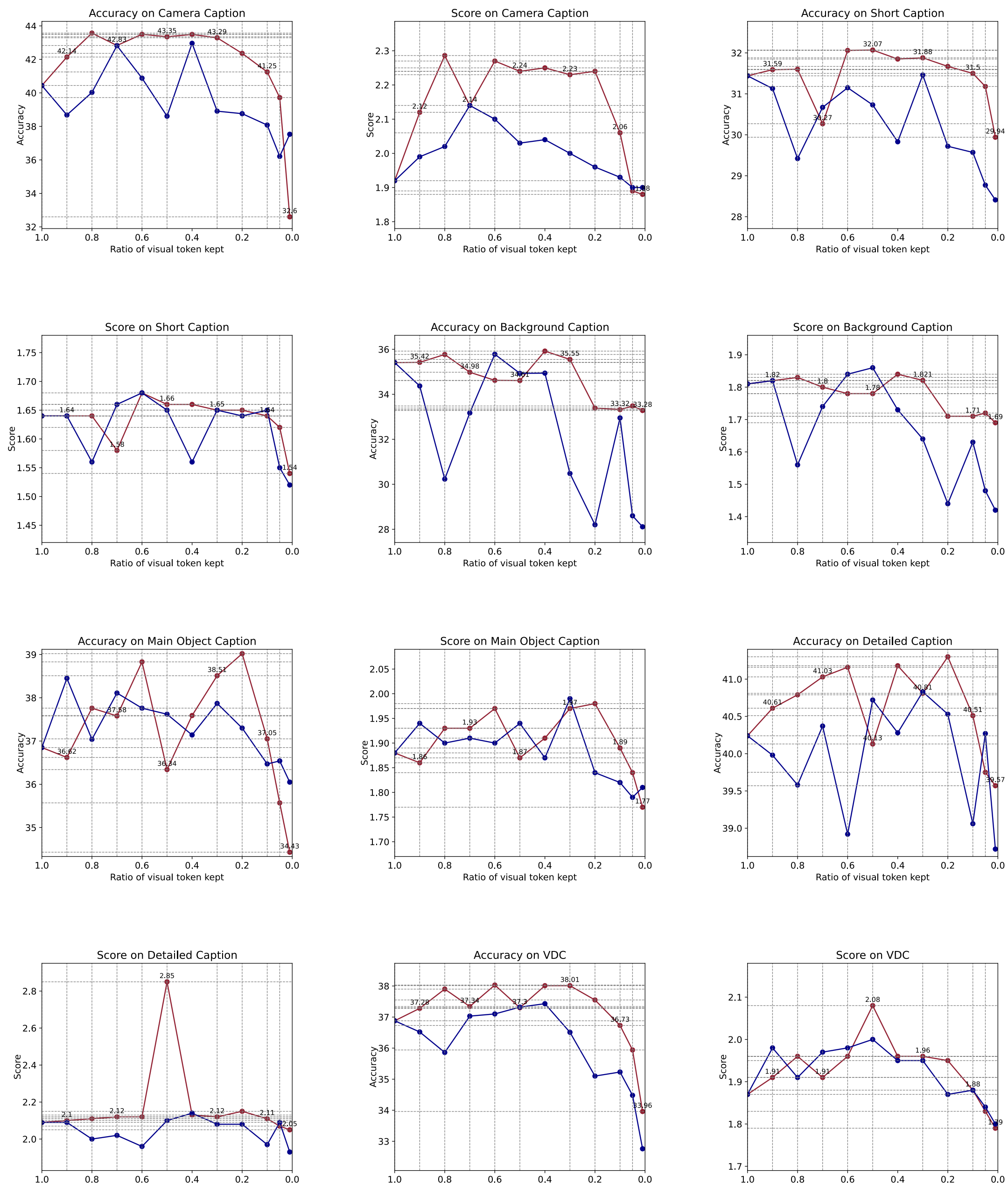


Figure D.6: Comparison between performance with and without the inclusion of full first-frame visual tokens during inference on VDC. As the visual token kept ratio varies, the blue curve indicates performance with slowfast inference, while the red curve represents performance without slowfast inference.

Appendix E
DETAILED TRAINING SETTINGS

We use CLIP ViT-H [1], Vicuna-1.5-7B [2] as the initialization of AuroraCap-7B. Training hyper-parameters for both stages are shown in Table E.4. For visual data preprocessing, we resize each image or video frame to the short side of 378 while keeping original aspect. Instead of doing center crop, we conduct bilinear position embedding interpolation for each training sample. For video data, we extract frames at 2 FPS uniformly. For token merging, we use constant schedule for each blocks.

[1] Huggingface Model: https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14-378

[2] Huggingface Model: https://huggingface.co/lmsys/vicuna-7b-v1.5-16k

Table E.1: Summary of datasets used for training AuroraCap in Pretraining stage.

Task	# Sample	Dataset
Image Captioning	1.3M	LAION-CC-SBU-595K [92], ShareGPT4V [22], ALLaVA-Caption-LAION-4V [18], ALLaVA-Caption-VFLAN-4V [18], DenseFusion [78]

Table E.2: Summary of datasets used for training AuroraCap in Vision stage. For classification, Reasoning, VQA, and Generation tasks, we adopt the dataset processed by M^3IT [76] to fit the training objective of language models.

Task	# Sample	Dataset
Captioning	1,925K	ShareGPT4V-PT [22], TextCaps [124], Image-Paragraph-Captioning [65]
Object-centric	438K	COST [58], ChatterBox [136], V* [148]
Classification	238K	COCO-GOI [88], COCO-Text [141], ImageNet [120], COCO-ITM [88], e-SNLI-VE [63], Mocheg [164], IQA [38]
Reasoning	100K	CLEVR [60], NLVR [129], VCR [168], VisualMRC [131], Winoground [135]
VQA	3,518K	VQA v2 [46], Shapes VQA [4], DocVQA [100], OK-VQA [98], Text-VQA [126], OCR-VQA [102], A-OK-VQA [122], ScienceQA [94] ST-VQA [10], ViQuAE [69], LLaVA-OneVision [70]
Generation	145K	Visual Storytelling [53], Visual Dialog [33], Multi30k [40]
Chinese	193K	COCO-Caption CN [80], Flickr-8k-Caption CN [79], multimodal Chat [181], FM-IQA [43], ChineseFoodNet [26]
Total	6.6M	For all datasets, we uniformly sample without duplication.

Table E.3: Summary of datasets used for training AuroraCap in Language stage.

Task	# Sample	Dataset
Image Captioning	1,779K	ShareGPT4V [22], ALLaVA-Caption-LAION-4V [18], ALLaVA-Caption-VFLAN-4V [18], DenseFusion [78], Face-Caption [31]
Video Captionin	1,659K	MiraData [61], LLaVA-Hound [173], ShareGPT4Video [23], Private Data
Image Instruction	9,742K	LLaVA-Mix-665K [90], LVIS-Instruct4V [145], ALLaVA-Instruct-LAION-4V [18], ALLaVA-Instruct-VFLAN-4V [18], Cambrian [137], M4-Instruct [91]
Video Instruction	268k	LLaVA-Hound [173], ShareGPT4Video [23]
Language-only	143K	Evol-Intruct-GPT4-Turbo-143K [18]
Total	15.0M	We duplicate video captioning and instruction datasets twice.

Table E.4: Training hyper-parameters for AuroraCap.

Hyper-parameters	Pretraining stage	Vision stage	Language stage
trainable parameters	MLP	ViT + MLP	ViT + MLP + LLM
warmup schedule	linear		
warmup start factor	1e-5		
warmup ratio	0.03		
learning rate schedule	cosine decay		
optimizer	AdamW [93]		
optimizer hyper-parameters	$\beta_1, \beta_2 = (0.9, 0.999)$		
weight decay	0.1		
max norm	1		
epoch	1		
peak learning rate	2e-4	1e-4	2e-5
total equivalent batch size	512	6,144	768
token keep proportion	100%	100%	10%

Appendix F

EVALUATION BENCHMARKS AND SETTINGS

We list all the hyper-parameters and prompt used for evaluation as shown in Table F.1. For our proposed VDC, we show the settings as following with the max number of the tokens of 1,024:

Type	Prompt used for evaluation
Camera	Describe any camera zooms, pans, or angle changes.
Background	Summarize the background setting of the video based on these frames.
Main Object	Describe the main subject, including their attributes and movements throughout the video.
Detail	Imagine the video from these frames and describe it in detail.

Table F.1: Evaluation settings summary for each benchmarks. For all benchmarks we set temperature, top p, number of beams to 0, 0, 1 respectively.

Benchmark	# Sample	# Tokens	Prompt
Flickr [113]	31,784	64	Provide a one-sentence caption for the provided image.
NoCaps [1]	4,500	64	Provide a one-sentence caption for the provided image.
COCO-Cap [88]	5,000	64	Provide a one-sentence caption for the provided image.
ChartQA [99]	2,500	16	Answer the question with a single word.
DocVQA [100]	5,349	32	Answer the question using a single word or phrase.
TextCaps [124]	3,166	64	Provide a one-sentence caption for the provided image.
GQA [54]	12,578	16	Answer the question using a single word or phrase.
POPE [84]	9,000	128	Answer the question using a single word or phrase.
MMMU [167]	900	16	Answer with the option letter from the given choices directly. Answer the question using a single word or phrase.
VQAv2 [46]	214,354	16	Answer the question using a single word or phrase.
MSR-VTT [154]	1,000	64	Provide a one-sentence caption for the provided video.
VATEX [146]	4,478	64	Provide a brief single-sentence caption for the last video below.
MSVD-QA [152]	1,161	64	Answer the question using a single word or phrase.
ActivityNet-QA [166]	8,000	64	Answer the question using a single word or phrase.
MSRVTT-QA [152]	6,513	64	Answer the question using a single word or phrase.
iVQA [160]	6,000	64	Answer the question using a single word or phrase.

Appendix G

LIMITATIONS

Table G.1: Comparison AuroraCap with LLM-based SoTA methods on visual question answering benchmarks under zero-shot setting.

Model	LLM	MMMU (900) Acc	GQA (12,578) Acc	POPE (9,000) F1	VQAv2 (214,354) Acc
LLaVA-1.5-7B	Vicuna-1.5-7B	35.30	61.97	85.87	76.64
LLaVA-1.5-13B	Vicuna-1.5-13B	34.80	63.24	85.92	78.26
LLaVA-1.6-7B	Vicuna-1.5-7B	35.10	64.23	86.40	80.06
LLaVA-1.6-13B	Vicuna-1.5-13B	35.90	65.36	86.26	80.92
AuroraCap-7B	Vicuna-1.5-7B	36.11	59.72	83.31	75.85

Table G.2: Limitation in terms of OCR capability compared with LLaVA models. Appendix F shows the introduction and metrics of each benchmark.

Model	LLM	ChartQA (2,500) Acc	DocVQA (5,349) Acc	TextCaps (3,166) Acc
LLaVA-1.5-7B	Vicuna-1.5-7B	18.24	28.08	98.15
LLaVA-1.5-13B	Vicuna-1.5-13B	18.20	30.29	103.92
LLaVA-1.6-7B	Vicuna-1.5-7B	54.84	74.35	71.79
LLaVA-1.6-13B	Vicuna-1.5-13B	62.20	77.45	67.39
AuroraCap-7B	Vicuna-1.5-7B	25.88	34.60	93.33

We evaluate AuroraCap on various visual question answering benchmarks as shown in Table G.1. Since the performance of the VQA task heavily depends on the performance of the LLM, we chose the same LLM for a fair comparison. Also, due the the limitation of OCR-related samples in training dataset, AuroraCap does not perform well in OCR at current stage as shown in Table G.2.

Appendix H

VDC GENERATION PROMPT TEMPLATE

Following [61], we utilize LLM to generate structured detailed captions. Given an input video, LLM return five detailed captions, including camera caption, short caption, background caption, main object caption and detailed caption for the entire video guided by our designed prompt template. The complete prompt is shown as followings:

Type	Prompt
SYSTEM	You are describing the video. Please provide detailed captions of the video from different aspects.
User	Please provide detailed and comprehensive captions for the following content: 1. Short Caption: Summarize the video in one detailed sentence, capturing key actions and the overall mood. 2. Background Caption: Provide a detailed description of the background, including objects, location, weather, time, and any dynamic elements such as movements in the environment. 3. Main Object Caption: Give a thorough description of the main subject's actions, attributes, interactions, and movements throughout the video frames, including changes in posture, expression, or speed. 4. Camera Caption: Describe the camera work in detail, including shot types, angles, movements, transitions, and any special effects used to enhance the video. 5. Reference Caption: Generate a detailed dense caption for the video that is at least 300 words long. The caption should capture all visible actions, environmental details, and the overall emotional atmosphere in depth. Describe in detail the interactions between the main subjects and their environment, including subtle nuances of their movements or expressions. Elaborate on the sounds, textures, and other sensory experiences depicted in the video. Discuss the camera techniques used extensively, including shot types, angles, movements, and transitions. Highlight the mood and tone of the video throughout, creating a rich narrative that connects viewers emotionally to the scene. Include comprehensive descriptions of background elements that add context and depth, such as weather conditions, time of day, and cultural or historical settings. Make sure to provide a vivid portrayal that is engaging, informative, and rich enough for AI to re-generate the video content.

No need to provide summary content. Do not describe each frame individually. Avoid using phrases like 'first frame'. The description should be rich enough for AI to re-generate the video. Please generate the response as a Python dictionary string with keys like 'short caption'. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string.

These are the frames from the video: [VIDEO KEYFRAME INPUT]

`GPT-4o` Short Caption: In a sunny outdoor setting, two workers engage in a collaborative effort to prepare a gravel base for a construction project, showcasing teamwork and focus.

Camera Caption: The camera captures the scene from a low angle, providing a unique perspective that emphasizes the workers' actions and the surrounding environment. The shots are primarily medium close-ups, focusing on the workers' upper bodies and hands as they interact. The camera occasionally shifts to a wider angle, revealing more of the workspace and background elements. There are smooth transitions between shots, maintaining a steady flow that enhances the viewer's engagement with the unfolding narrative. The use of a slightly fish-eye lens effect adds depth to the scene, making the environment feel more immersive.

Background Caption: The scene is set in a residential area with a well-maintained lawn and a brick wall partially constructed in the background. The ground is covered with a white plastic sheet, and scattered gravel is visible. A few wooden planks lie on the ground, and tools are scattered around, including a black water bottle and a power tool.

Main Object Caption: The main subjects are two workers, one wearing a gray sweatshirt and gloves, and the other dressed in casual work attire. They are actively engaged in conversation, with the first worker gesturing with his hands to emphasize points while explaining the task at hand. His posture is relaxed yet focused, indicating a collaborative atmosphere. The second worker listens attentively, occasionally nodding and responding with gestures, showcasing a dynamic interaction. Both workers exhibit a sense of camaraderie and professionalism as they prepare for the next steps in their project.

Reference Caption: The video unfolds . . . (300 words) . . . through them.

User The video has been describe from the following aspects:1. short caption: short caption, 2. background caption: background caption, 3. main object caption: main object caption, 4. camera caption: camera caption, 5.Reference Caption: reference caption Please generate a detailed dense caption for the video that is pretty long. You should expand the reference caption upon the information provided in the short caption, background caption, main object caption, and camera caption. Ensure that the detailed caption does not introduce any new entities or relationships that were not mentioned in the previous captions. Make sure to provide a vivid portrayal that is engaging, informative, and rich enough for AI to re-generate the video content. "Avoid using phrases like 'first frame', 'short caption', 'background caption', 'main object caption', and 'camera caption'. The description should be rich enough for AI to re-generate the video.

`GPT-4o` Detailed Caption: The video opens with ... (500 words) ... literature that surrounds them.

User [Manually check and refine]

Appendix I

QUESTION-ANSWER PAIRS GENERATION PROMPT TEMPLATE OF VDCscore

To decompose the ground-truth structured detailed captions in VDC, we utilize `GPT-4o` as the LLM assistant to generate numerous short question-answer pairs for subsequent evaluation. The full prompt and example cases are presented as followings:

Type	**Prompt**
SYSTEM	You are an intelligent chatbot designed for generating 20 question-answer pairs given a detailed description of a video or image. You are describing the video. Here's how you can accomplish the task: INSTRUCTIONS: - Cover the main objects and actions in the video or image. - The questions should be open-ended and start with 'What', 'Who', 'Where', 'When', 'Why', 'How', etc. - The answer should be a short sentence or phrase. - Generate 20 question-answer pairs.
User	Please generate 20 question-answer pairs given a detailed description of a video or image: detailed description: The camera work is dynamic, featuring a mix of close-up shots and wider angles that provide context to the environment. The shots are slightly shaky, adding a sense of realism and immersion to the viewer. Transitions between shots are smooth, with the camera often panning or tilting to follow the subject's movements. The use of low light creates a moody atmosphere, with shadows enhancing the visual interest of the scene ... Please generate the response in the form of a Python list of tuple with the question and the corresponding answer. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python list of tuple. For example, your response should look like this: [(the question 1, the answer 1), (the question 2, the answer 2), ...].
`GPT-4o`	["question": "What is the style of camera work in the video?", "answer": "dynamic", "question": "What is the effect of shaky shots in the video?", "answer": "realism and immersion", "question": "What type of transitions are used in the video?", "answer": "smooth", "question": "Why is the moody atmosphere created in the video?", "answer":

"through low light", "question": "What type of camera movement is achieved through panning and tilting?", "answer": "following the subject's movements", "question": "What is the effect of the camera movement on the viewer?", "answer": "sense of realism and immersion", . . .]

User [Manually check and refine]

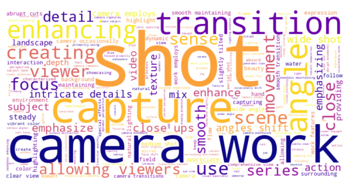
Appendix J

MORE STATISTICS INFORMATION OF VDC

Figure J.1 indicates the word distribution of the structured captions in VDC. Table J.1 illustrates the distribution of the question categories we provided for VDCscore calculation.

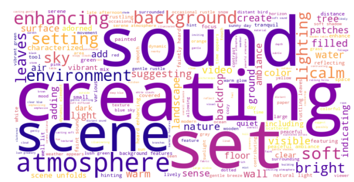
Table J.1: Distribution of the question categories in VDCscore.

Type	Background	Camera	Short	Main object	Detailed
Environment & Scene	3,464	2,543	4,401	3,097	3,549
Character & Object	6,272	3,779	4,834	6,035	6,440
Intent & Outcomes	4,585	5,105	5,067	4,474	4,640
Action	2,861	1,198	1,857	3,192	1,932
Attributes & Relationship	944	334	1,388	1,478	1,280



(a) Camera Caption

(b) Short Caption



(c) Background Caption

(d) Main Object Caption

(e) Detailed Caption

(f) Total

Figure J.1: Word cloud of different structured captions in VDC, showing the diversity.

Appendix K
MORE STATISTICS INFORMATION OF VDCscore

We generate a total of 96,902 question-answer pairs for VDC, with an average of 18.87 pairs per detailed caption. As depicted in Figure K.1, each section of the structured captions includes a similar number of question-answer pairs. Additionally, Figure K.2 presents the distribution of question types generated for VDC. To enhance the evaluation of detailed captions, we configure all questions as open-ended. `Environment & Scene` encompasses inquiries about location, environment, atmosphere, scene, and time. `Character & Object` focuses on entities within the caption, such as people, animals, plants, and objects, while `Attribute & Relation` examines their attributes and interrelations. `Intent & Outcomes` addresses deeper interpretative questions regarding methods, purposes, reasons, and outcomes. We further analyze the distribution of generated question types within structured captions. Main object captions predominantly feature `Character & Object` questions, whereas `Environment & Scene` questions are more prominent in background captions, and `Camera` questions constitute a larger proportion in camera.

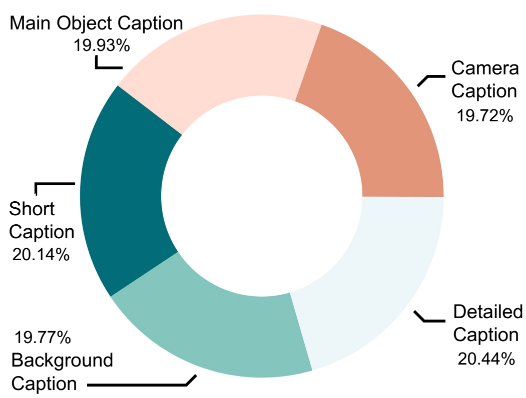


Figure K.1: Question-answer pairs proportaion in structured captions.

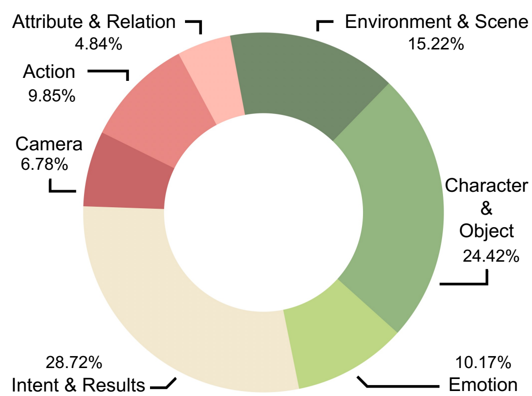


Figure K.2: Distribution of question type in VDCScore.

Appendix L

CALCULATION OF ELO RANKING

In this section, we present the methodology for evaluating and ranking AuroraCapand various models using the Elo rating system as shown in Figure L.1. The parameters used in the simulation are summarized in Table L.1.

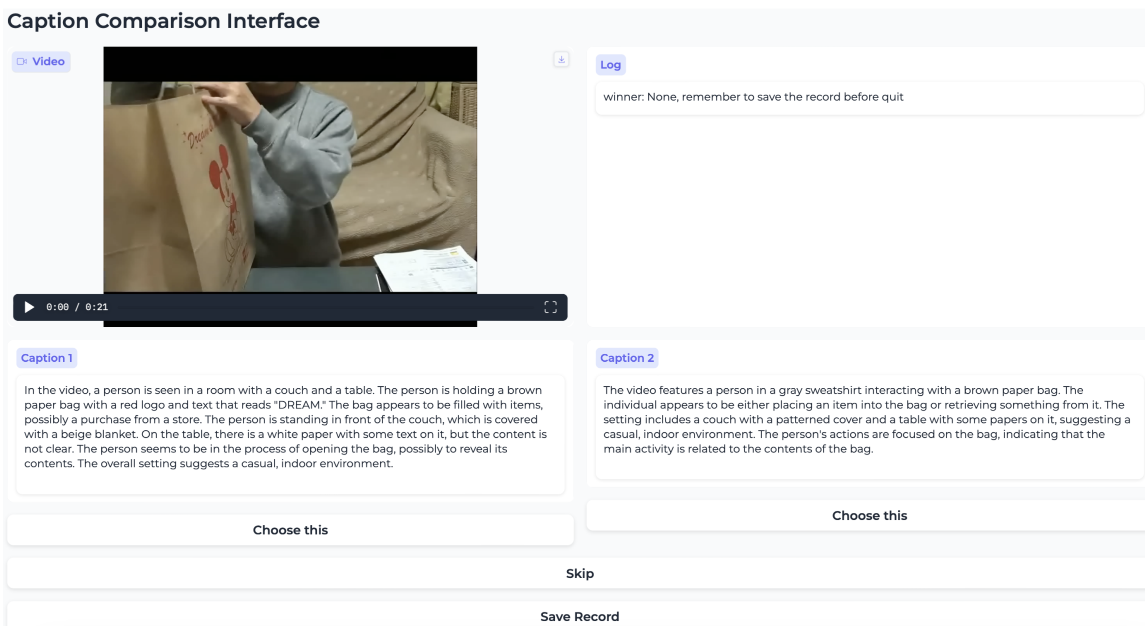


Figure L.1: GUI screen for Elo ranking.

Table L.1: Elo parameter setting.

Parameter	Number
initial Elo mean	1,000
Elo standard deviation	300
base of logarithm	10
scaling factor	400
K-factor	32
minimum Elo rating	700
number of simulated matches	2,778

Appendix M
CASE STUDY

We perform an extensive case study of AuroraCap on a variety of videos for video detailed captioning. As shown as followings, AuroraCap is capable of providing excellent detailed captions regarding the camera motion, background and main object with less hallucination.

Example	Caption
Figure M.1	[short caption] A man with a mustache and a woman with a ponytail are sitting at a table.
Figure M.2	[short caption] A man plays the drums in a recording studio.
Figure M.3	[detailed caption] The video depicts a serene beach scene where a young woman stands on the sandy shore, gazing out towards the ocean. She is wearing a black beanie and a pink jacket, adding a pop of color to the otherwise muted scene. The beach, a vast expanse of sand, stretches out in front of her, meeting the ocean at the horizon. The ocean, a vast body of water, is visible in the background. The beach is bathed in a soft, diffused light, creating a dreamy atmosphere. The girl's gaze is directed towards the horizon, suggesting a sense of wonder or contemplation. The image is slightly blurred, adding a dreamy quality to the scene. The woman's position on the beach, coupled with the gentle waves of the ocean, suggests a moment of contemplation or admiration. The relative positions of the objects suggest a peaceful day at the beach, with the girl possibly enjoying the serene view of the ocean. The colors are mostly muted, with the girl's pink jacket standing out against the sandy beach and the blue ocean. The blurred background and the out-of-focus elements, such as the ocean and the sky, contribute to the sense of tranquility and focus on the woman. There is no text present in the video, and the colors are muted, with the exception of the pink jacket, which stands out against the more subdued tones of the surroundings.
Figure M.4	[detailed caption] The video features a man with a beard and long hair, lying on the ground with his head resting on a tree trunk. He is wearing a colorful shirt with a mix of

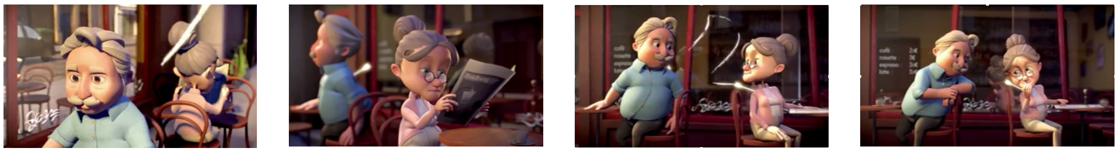


Figure M.1: Video example of MSR-VTT [154] benchmark.

Figure M.2: Video example of VATEX [146] benchmark.

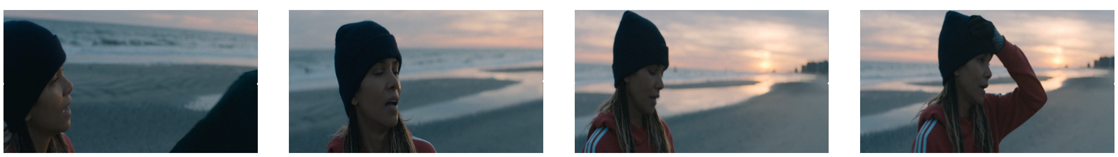


Figure M.3: Video example.

Figure M.4: Video example.

Figure M.5: Video example.

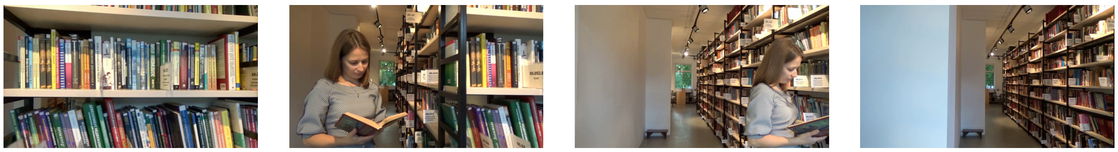


Figure M.6: Video example.

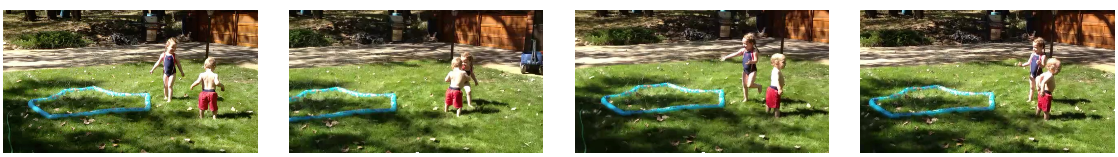


Figure M.7: Video example.

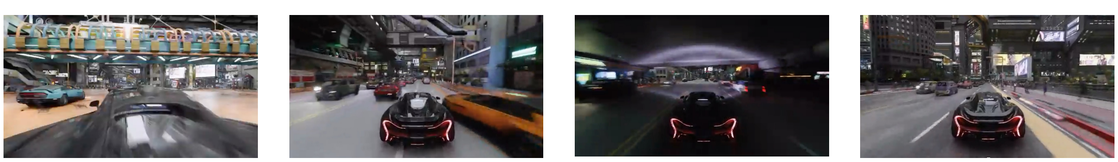


Figure M.8: Video example.

orange and green patterns. The man's face is partially obscured by the tree trunk, but his eyes are visible, looking upwards. He appears to be smiling or laughing, with his mouth open. The man's arms are crossed over his chest, suggesting a relaxed or playful posture. The background is blurred, but it appears to be a natural outdoor setting, possibly a forest or park. In the background, another person is partially visible, their presence subtly hinted at by a hand reaching out from the left side of the frame. The colors in the video are warm, with the man's shirt standing out against the muted background. The man's hair and beard are dark, and he has a light complexion. The overall mood of the video is lighthearted and carefree. There are no visible texts or other objects in the video, and the relative positions of the objects remain constant with the man in the foreground and the second person in the background.

Figure M.5 [background caption] The background features a race track with visible tire marks and barriers, surrounded by grassy areas and a few scattered trees. The track is set in a rural or semi-rural location, with hills in the distance and a cloudy sky overhead, suggesting overcast weather conditions. The track itself appears well-maintained with a smooth surface, designed for high-speed racing. The weather, indicated by the cloudy sky, contributes to the overall ambiance of the scene, enhancing the sense of a cool, possibly early morning or late afternoon setting.

Figure M.6 [camera caption] The camera maintains a steady, frontal view throughout the video, capturing the woman's expressions and the intricate details of the bookshelves. The camera starts at the right edge of the bookshelf, moving across the shelf, and ends at the left edge of the bookshelf. The framing is consistent, focusing on the woman and the bookshelves, with the camera positioned at a medium distance to the subject. This movement provides a comprehensive view of the bookshelf, showcasing the variety of books and their arrangement on the shelves. The camera occasionally pans to reveal the depth of the library, showcasing the rows of books and the inviting atmosphere. The use of natural light enhances the visual appeal, creating a warm and inviting tone throughout the video.

Figure M.7 [camera caption] The view shot remains relatively static, focusing on the children playing in the backyard. The camera angle is at eye level, capturing the scene from a distance that allows both children to be visible. There is minimal camera movement, maintaining a steady focus on the children and their activities. The sequence of video frames suggests a continuous moment of play without significant changes in shooting angles or camera movement.

Figure M.8 [main object caption] The main subject in the video, a black car, is seen driving down a street that appears to be in a state of disarray. The car moves steadily forward, navigating around obstacles such as a blue car parked on the side of the road. The car's movement is smooth and continuous, suggesting it is either in motion or has just come to a stop. The environment around the car is chaotic, with debris scattered across the road and signs of destruction, indicating a recent event or disaster. The car's position remains central in the frame, with the camera angle focused on it from a slightly elevated perspective, possibly from a vehicle or a structure above.