VIDCOMPOSITION: Can MLLMs Analyze Compositions in Compiled Videos?

Yunlong Tang[1,*], Junjia Guo[1,*], Hang Hua[1], Susan Liang[1], Mingqian Feng[1], Xinyang Li[1], Rui Mao[1], Chao Huang[1], Jing Bi[1], Zeliang Zhang[1], Pooyan Fazli[2], Chenliang Xu[1†]

[1]University of Rochester, [2]Arizona State University

{yunlong.tang, mingqian.feng, jing.bi, chenliang.xu}@rochester.edu, pooyan@asu.edu
{jguo40, xli190, rmao6, zzh136}@ur.rochester.edu, {hhua2, sliang22, chuang65}@cs.rochester.edu

Abstract

The advancement of Multimodal Large Language Models (MLLMs) has enabled significant progress in multimodal understanding, expanding their capacity to analyze video content. However, existing evaluation benchmarks for MLLMs primarily focus on abstract video comprehension, lacking a detailed assessment of their ability to understand video compositions, the nuanced interpretation of how visual elements combine and interact within highly compiled video contexts. We introduce VidComposition, a new benchmark specifically designed to evaluate the video composition understanding capabilities of MLLMs using carefully curated compiled videos and cinematic-level annotations. VidComposition includes 982 videos with 1706 multiple-choice questions, covering various compositional aspects such as camera movement, angle, shot size, narrative structure, character actions and emotions, etc. Our comprehensive evaluation of 33 open-source and proprietary MLLMs reveals a significant performance gap between human and model capabilities. This highlights the limitations of current MLLMs in understanding complex, compiled video compositions and offers insights into areas for further improvement. The leaderboard and evaluation code are available at https://yunlong10.github.io/VidComposition/.

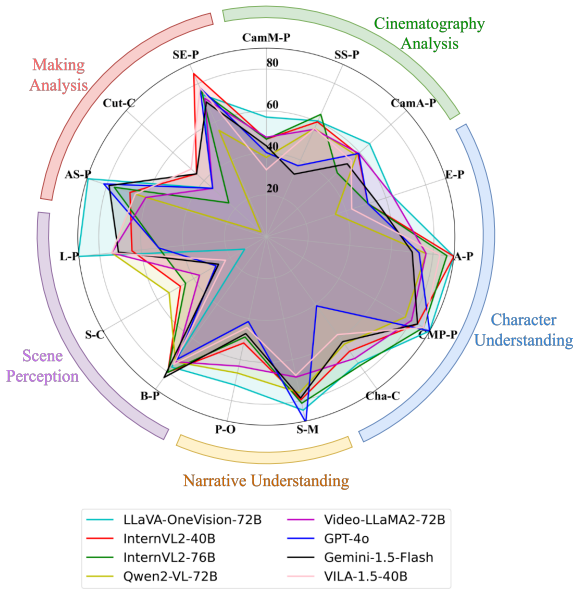

Figure 1. Top MLLMs' performance on VIDCOMPOSITION, across 15 tasks of 5 aspects of video composition understanding: Cinematography Analysis, Character Understanding, Narrative Understanding, Scene Perception, and Making Analysis.

1. Introduction

Recent advancements in Multimodal Large Language Models (MLLMs) [1, 3, 8, 15, 27, 43, 47] have greatly enhanced capabilities in understanding multimodality. However, while current benchmarks [10, 12, 25, 33] for evaluating MLLMs assess general image or video comprehension, they lack a detailed focus on video composition, the nuanced interpretation of how visual elements combine and interact within compiled videos. Compiled videos refer to those created by editing and integrating multiple clips, scenes, or sequences, either from various sources or from different segments of a single recording, *e.g.* films, TV series, documentaries, animations, vlogs, *etc*. These videos are carefully constructed to create a seamless flow and include richer compositions, requiring shot-by-shot analysis to interpret.

Shot-by-shot analysis, a technique where creators meticulously break down the elements of a video, serves as a vital tool for understanding video composition in depth. This level of understanding, essential in film analysis and video production, goes beyond general scene or action recognition, requiring an in-depth grasp of compositional elements such

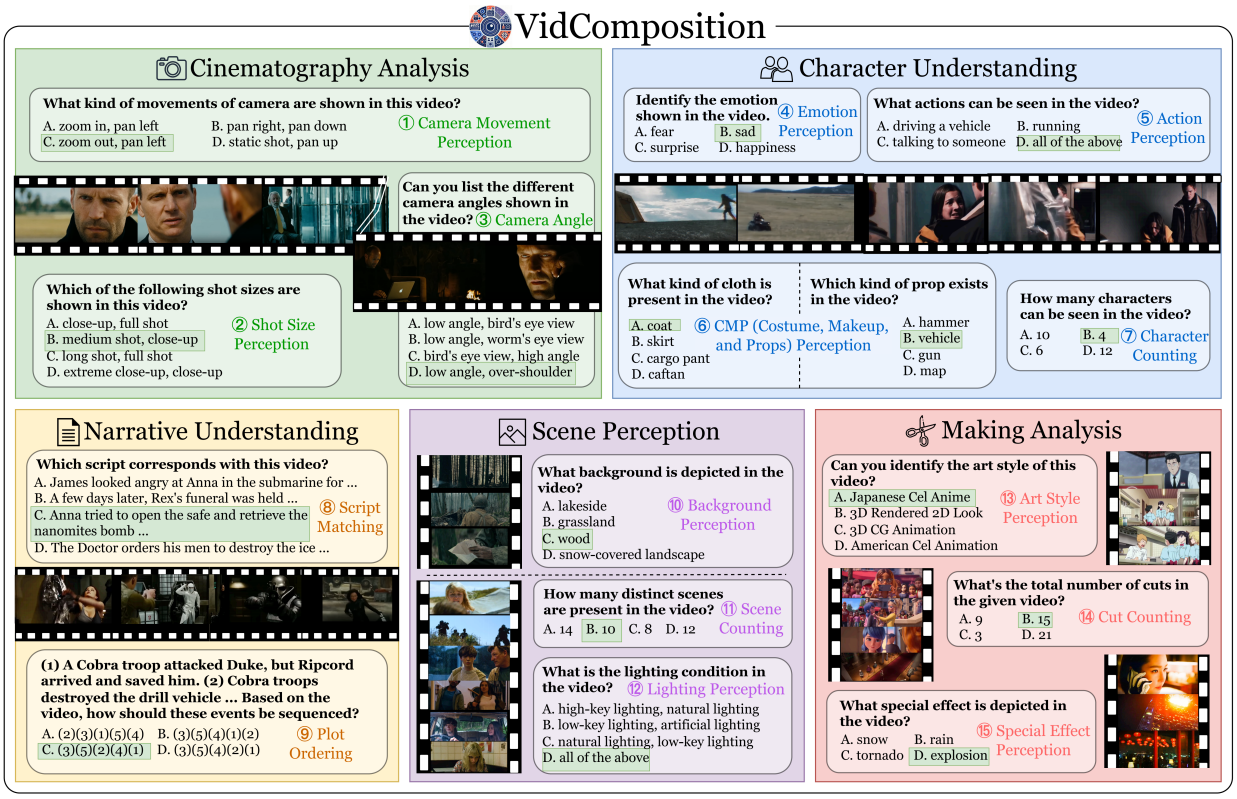

Figure 2. VIDCOMPOSITION comprises 15 categories of high-quality QA pairs, focusing on five aspects of compositions in compiled videos: cinematography, character, narrative, scene, and making. The correct answers are highlighted.

as camera movements, shot sizes, narrative structures, and character dynamics. This analysis also captures the intricate layers of visual storytelling by deconstructing how technical and artistic choices shape the viewing experience. However, achieving this fine-grained level of composition understanding remains a significant challenge for existing MLLMs, which primarily operate on broader, more coarse-grained interpretations of video content.

Through investigating existing benchmarks, we identified their limitations in evaluating MLLM in video composition understanding. As shown in Table 1, the benchmarks in the first group [10, 16, 31] primarily focus on static images and overlook the dynamic aspects of visual content. Among these, Winoground [48] and MMComposition [16] attempts to assess the compositionality of MLLMs, though it is limited to image-based evaluations. The second group consists of traditional video benchmarks [21, 53, 54, 58], which are less effective at addressing the specific limitations of modern MLLMs. While TVQA [21] includes a compositional video QA component, its compositionality is relatively coarse-grained, limited to basic question types like "who," "when," "where," "how," and "what." The third group highlights recent benchmarks [5, 12, 25, 32, 40, 44, 50] developed to assess MLLMs' video comprehension capabilities. Although these benchmarks incorporate tasks that touch on compositional understanding, their evaluations of compositionality remain limited. Additionally, most videos in these benchmarks are natural-shot rather than compiled, posing a challenge for models trained on natural footage to effectively interpret the increasingly prevalent edited and compiled videos seen on modern online video platforms.

Recognizing the gap in existing evaluation methods, we introduce VIDCOMPOSITION, a new benchmark designed to assess MLLMs on understanding video composition at a cinematic level. VIDCOMPOSITION includes 982 carefully curated videos and 1,706 multiple-choice questions, featuring meticulously annotated clips from films, TV series, animations, commentary videos, *etc*. These questions include five key areas of video composition: Cinematography Analysis, Character Understanding, Narrative Understanding, Scene Perception, and Making Analysis, spanning 15 distinct tasks. Each area captures critical aspects of compositional understanding, *e.g*. camera movements, angles, shot sizes, narrative structures, characters, scenes, cuts, special effects, *etc*., providing a extensive framework for evaluating the nuanced comprehension required in cinematic contexts.

We evaluate 33 state-of-the-art MLLMs on VIDCOMPOSITION, including 27 open-source and 6 proprietary models, revealing a substantial performance gap between MLLMs and human-level comprehension in video composition un-

Table 1. A comparative overview of various benchmarks across several dimensions, such as data format (image **I** or video **V**), the size of dataset *for evaluation* (**#Data**), the number of tasks covered (**#Task**), whether the dataset supports compositional question answering (**Compositional QA**), the presence of **Compiled Videos** and **Fine-Grained** sub-tasks, and the annotation method (manual or automatic/manual, indicated by **Anno.**).

Benchmark	I/V	#Data	#Task	Compositional QA	Compiled Videos	Fine-Grained	Anno.
Winoground [48]	I	400	8	✓	-	✗	M
MME [10]	I	1.1k	14	✗	-	✓	M
MMBench [33]	I	1.7k	20	✗	-	✓	A+M
MMComposition [16]	I	4.3k	13	✓	-	✓	M
MSVD-QA [54]	V	504	5	✗	✗	✗	A
MSRVTT-QA [54]	V	2.9k	5	✗	✗	✗	A
TGIF-QA [18]	V	9.6k	4	✗	✗	✗	A
TVQA [21]	V	2.2k	8	✓	✓	✗	A+M
ActivityNet-QA [58]	V	5.8k	4	✗	✗	✗	M
NExT-QA [53]	V	1k	8	✗	✗	✗	A
AutoEval-Video [5]	V	327	9	✗	✗	✗	A+M
Video-Bench [40]	V	5.9k	10	✗	✗	✗	A+M
LVBench [50]	V	500	6	✗	✗	✗	M
MVBench [25]	V	3.6k	20	✗	✗	✓	A+M
MovieChat-1k [44]	V	100	8	✗	✓	✓	M
TempCompass [32]	V	410	4	✗	✗	✓	M
Video-MME [12]	V	900	12	✗	✗	✓	M
VIDCOMPOSITION	V	982	15	✓	✓	✓	M

derstanding. As shown in Figure 1, although top models [1, 7, 8, 23, 28, 43, 49] perform well on basic perception tasks (*e.g.* action perception), they fall short in comprehending complex video compositions, particularly in cinematography. This performance disparity underscores current models' limitations in capturing intricate, multi-layered video structures. Additional experiments further analyze factors influencing MLLM performance, such as the number of frames provided as input, the resolution of visual encoders, the size of language decoders, and the data volume for fine-tuning, yielding insights for future advancements in model design. Overall, our benchmark offers valuable insights for enhancing MLLMs and also suggests applications in video generation where MLLMs could assist in automatically evaluating the compositional quality of generated videos.

In summary, our contribution is three-fold:

- We introduce VIDCOMPOSITION, a novel, human-annotated, high-quality benchmark for evaluating fine-grained video composition understanding in MLLMs.
- We comprehensively evaluate 33 MLLMs for video understanding with VIDCOMPOSITION. The results show the challenging nature of VIDCOMPOSITION and the **substantial gap** between MLLMs' and humans' capabilities in video composition understanding.
- We analyze the critical factors that influence the performance of MLLMs systematically, providing potential directions for model improvement and future advancements.

2. VIDCOMPOSITION

2.1. Overview and Terminology

VIDCOMPOSITION contains 982 compiled videos with 1706 human-annotated multiple-choice questions for video composition understanding, including 5 main categories: Cinematography Analysis (**CA**), Character Understanding (**CU**), Narrative Understanding (**NU**), Scene Perception (**SP**), and Making Analysis (**MA**); and 15 sub-tasks: Camera Movement Perception (**CamM-P**), Shot Size Perception (**SS-P**), Camera Angle Perception (**CamA-P**), Emotion Perception (**E-P**), Action Perception (**A-P**), Costume, Makeup and Props Perception (**CMP-P**), Character Counting (**Cha-C**), Script Matching (**S-M**), Plot Ordering (**P-O**), Background Perception (**B-P**), Scene Counting (**S-C**), Lighting Perception (**L-P**), Art Style Perception (**AS-P**), Cut Counting (**Cut-C**), and Special Effect Perception (**SE-P**). Examples of each task can be found in Figure 2. The detailed definitions of each task are provided in Supplementary.

2.2. Dataset Curation Process

Video Collection and Filtering. Our dataset comprises videos sourced from the Internet, focusing on compiled videos primarily derived from commentary videos for movies, TV series, and animations, which have no copyright concerns. These videos typically include subtitles and scripts uploaded by users, which assist in later annotation stages. We further refine the collected videos by filtering out inappropriate content, such as clips that may cause psychological distress or those flagged as sensitive by API-based models. The average duration of the collected videos is about 20 minutes. For videos with subtitles or scripts, we extract the timestamps marking the start and end of each subtitle. With this information, we further segment the video into coherent sections whose average length is 794 frames. To avoid models from predicting answers relying on speech, we removed the audio of the videos.

Human Annotation. To ensure the quality and reliability of the dataset, we engage multiple human annotators, assigning each video segment to several annotators to minimize potential biases. All questions are meticulously designed to address specific tasks. For perception tasks such as **A-P**, **E-P**, **CMP-P**, **B-P**, and for counting tasks such as **Cha-C**, **S-C**, and **Cut-C**, annotators watch the video segment and write the correct answer alongside several incorrect (distractor) options. For tasks such as **CamM-P**, **SS-P**, **CamA-P**, **L-P**, **SE-P**, and **AS-P**, we provide a predefined set of selectable labels. For example, for **CamM-P**, the labels include *zoom in*, *zoom out*, *pan left*, *pan right*, *pan up*, *pan down*, and *static shot*. Annotators choose appropriate labels for each video segment, which are then used as correct options, while distractors are randomly selected from the remaining labels, ensuring they differ from the correct options. For **S-M**, we

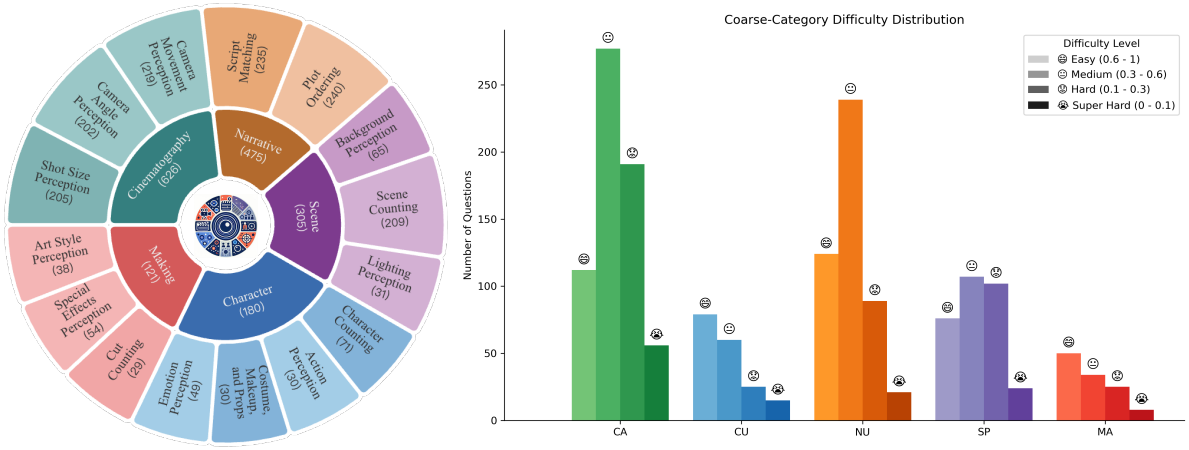

Figure 3. (**Left**) Task statistics in VIDCOMPOSITION, organized into five main categories: Cinematography Analysis (**CA**), Character Understanding (**CU**), Narrative Understanding (**NU**), Scene Perception (**SP**), and Making Analysis (**MA**), comprising a total of 15 sub-tasks. The number of QA pairs is shown in parentheses below each task. (**Right**) The difficulty distribution across these five categories. If a question is answered correctly by more than 60% of MLLMs, it will be labeled as "Easy." Conversely, if a question is answered correctly by fewer than 10% of MLLMs, it will be labeled as "Super Hard."

use the video's commentary script, extracted from the subtitle file, as the correct option, with distractor options sourced from nearby segments' scripts to create plausible alternatives. For **P-O**, we segment the commentary script into multiple parts, shuffling and inserting them into the question with sequence numbers. The correct answer is the original order of the script, while other options are generated by randomizing these sequence numbers.

Quality Control. To ensure the quality of our benchmark, each video and corresponding QA pair undergoes multiple rounds of review. We implemented an annotation review system (see the user interface in Supplementary), which displays the annotated video, question, and answer options alongside an additional feedback option for reviewers to provide corrections or comments. Reviewers are required to attempt each question themselves; if they identify errors in the question or options, they can select the feedback option to either suggest improvements to the question or specify what they believe is the correct answer. After each round of review, all feedback submitted through the annotation review system is analyzed and used to enhance the annotation quality further. This iterative quality control process ensures accuracy and consistency across annotations, minimizes errors, and refines question clarity for each task.

2.3. Evaluation Metrics

To obtain model predictions, each question is structured within a predefined prompt template $\mathcal{P}$ that includes the question text and associated options $\mathcal{O}_q$ (A, B, C, D, along with descriptive texts). This prompt $\mathcal{I}_q$ is then fed into the model $\mathcal{M}$, which is expected to output a single character representing its predicted answer (one of $\{A, B, C, D\}$). The prediction process is formalized in Algorithm 1. The prompt template $\mathcal{P}$ can be found in Supplementary.

Algorithm 1 Model Prediction

Input: Question q, Options $\mathcal{O}_q$
Output: Prediction $\mathcal{A}$
$\mathcal{I}_q \leftarrow \mathcal{P}(q, \mathcal{O}_q)$
$\mathcal{R}_q \leftarrow \mathcal{M}(\mathcal{I}_q)$
$$\mathcal{A}_q \leftarrow \begin{cases} \mathcal{R}_q & \text{if } \mathcal{R}_q \in \{A, B, C, D\}, \\ AM(\mathcal{R}_q) & \text{if more letters in } \mathcal{R}_q, \\ RS(\{A, B, C, D\}) & \text{otherwise.} \end{cases}$$

The model output $\mathcal{R}_q$ is first checked for validity as a single character from $\{A, B, C, D\}$. An *Answer-Matching (AM)* function identifies a valid option if the output includes multiple characters. The implementation of the *AM* function can be found in the Supplementary Materials. A *random selection (RS)* function from $\{A, B, C, D\}$ is used to generate $\mathcal{A}_q$ if no valid match is found.

Once $\mathcal{A}_q$ is obtained, we evaluate its accuracy based on whether it matches the correct answer for each question. Let $\mathcal{S} = \{\mathcal{S}_i = \{\mathcal{Q}_j\}_{j=1}^{N_i}\}_{i=1}^{|\mathcal{S}|}$ represent our dataset, where each sub-task $\mathcal{S}_i$ contains a set of questions $\mathcal{Q}_j$ across a total of $|\mathcal{S}|$ sub-tasks. For each question $q \in \mathcal{S}$, let $\mathcal{G}_q$ represent the correct answer, and $\mathcal{A}_q$ represent the model's answer. The score for question q, denoted as s_q, is calculated as follows:

$$s_q = \begin{cases} 1, & \text{if } \mathcal{A}_q = \mathcal{G}_q, \\ 0, & \text{if } \mathcal{A}_q \neq \mathcal{G}_q. \end{cases} \quad (1)$$

A score of 1 is assigned if the model's prediction $\mathcal{A}_q$ matches the correct answer $\mathcal{G}_q$; otherwise, the score is 0. Each sub-task's accuracy ACC_i is calculated as the average score across its N_i questions: $\text{ACC}_i = {N_i}^{-1} \sum_{j=1}^{N_i} s_{q_j}$, where N_i is the total number of questions in sub-task $\mathcal{S}_i$. The overall accuracy is computed as the ratio of total correct answers

Table 2. The comprehensive evaluation of 30 MLLMs on VIDCOMPOSITION, including open source models and API-based models. The **best** results are in bold, and second best results are in underlined, respectively.

Method	Cinematography Analysis			Character Understanding				Narrative Underst.		Scene Perception			Making Analysis			Overall
	CamM-P	SS-P	CamA-P	E-P	A-P	CMP-P	Cha-C	S-M	P-O	B-P	S-C	L-P	AS-P	Cut-C	SE-P	
Human	84.1	85.4	80.0	82.6	92.3	92.9	94.1	97.0	97.5	94.4	80.2	81.8	85.7	87.5	94.7	86.26
LLaVA-OneVision-72B [23]	**57.1**	60.5	**66.3**	**63.3**	**90.0**	**90.0**	74.6	84.7	**72.4**	76.9	12.0	**90.3**	**89.5**	34.5	74.1	**63.31**
InternVL2-40B [6]	46.6	60.0	58.9	51.0	**90.0**	83.3	67.6	79.6	51.9	80.0	47.4	64.5	68.4	44.8	**85.2**	60.73
InternVL2-76B [6]	46.6	**63.9**	45.5	51.0	86.7	86.7	**76.1**	81.3	49.0	80.0	44.5	51.6	76.3	24.1	75.9	58.73
Qwen2-VL-72B [49]	37.9	56.6	57.9	34.7	76.7	76.7	63.4	76.2	66.5	73.8	53.6	74.2	65.8	3.4	55.6	58.68
Video-LLaMA2-72B [8]	47.5	56.1	59.4	**63.3**	76.7	80.0	71.8	68.5	63.2	73.8	36.8	74.2	60.5	34.5	72.2	58.62
InternVL2-8B [6]	55.3	56.6	59.4	44.9	80.0	83.3	59.2	67.2	40.6	78.5	32.5	64.5	52.6	31.0	72.2	54.63
GPT-4o [1]	40.2	37.1	59.4	51.0	73.3	**90.0**	40.8	**90.6**	41.4	72.3	27.8	51.6	81.6	34.5	77.8	52.93
Gemini-1.5-Flash [43]	43.4	32.7	52.0	55.1	70.0	48.4	62.0	78.7	47.3	**83.1**	26.3	71.0	78.9	44.8	70.4	52.40
VILA-1.5-40B [28]	32.0	56.6	54.5	42.9	83.3	86.7	57.7	67.7	44.4	75.4	22.5	74.2	65.8	48.3	77.8	51.23
GPT-4o mini [1]	33.8	49.8	50.5	49.0	80.0	**90.0**	31.0	79.6	41.4	66.2	26.8	61.3	76.3	20.7	79.6	50.23
Gemini-1.5-Pro [43]	33.8	51.7	51.5	51.0	73.3	80.0	70.4	47.7	36.4	73.8	37.3	71.0	84.2	**58.6**	75.9	49.36
Qwen2-VL-7B [49]	20.1	46.8	37.1	38.8	70.0	76.7	54.9	73.2	49.4	72.3	52.2	61.3	42.1	17.2	70.4	49.30
Oryx-7B [34]	34.7	54.1	57.4	57.1	80.0	73.3	66.2	48.5	34.7	73.8	41.6	61.3	39.5	20.7	66.7	48.77
Gemini-1.5-Flash-8B [43]	43.4	45.9	56.9	36.7	70.0	76.7	35.2	69.8	36.0	73.8	26.8	48.4	71.1	24.1	64.8	48.59
Video-LLaMA2.1 [8]	44.3	35.6	39.6	51.0	76.7	83.3	50.7	60.9	45.2	75.4	35.4	58.1	36.8	20.7	81.5	47.77
VideoChat2 [25]	24.2	58.0	42.1	44.9	66.7	83.3	60.6	62.6	27.6	73.8	**55.0**	35.5	50.0	10.3	59.3	47.36
InternVL2-26B [6]	33.3	47.8	39.6	55.1	76.7	83.3	56.3	68.9	33.9	76.9	25.4	45.2	34.2	41.4	75.9	46.42
LongVA [61]	24.7	41.0	48.0	40.8	70.0	73.3	42.3	51.9	32.2	72.3	42.6	48.4	52.6	34.5	70.4	43.73
MiniCPM-V2.6 [55]	28.3	43.4	43.6	53.1	73.3	80.0	50.7	59.1	23.0	75.4	22.0	71.0	57.9	20.7	72.2	42.49
InternVL2-4B [6]	27.4	42.9	26.2	32.7	66.7	73.3	49.3	60.4	28.0	78.5	41.6	35.5	44.7	10.3	72.2	41.68
Video-LLaMA2.1-AV [8]	27.4	45.9	38.6	55.1	73.3	76.7	47.9	46.4	30.1	81.5	25.8	45.2	34.2	34.5	83.3	41.50
VILA-1.5-8B [28]	31.5	40.0	37.6	51.0	63.3	66.7	40.8	40.9	26.8	70.8	37.8	41.9	60.5	44.8	59.3	40.21
GPT-4-turbo [1]	23.7	37.1	35.1	46.9	63.3	80.0	25.4	54.9	36.4	50.8	29.7	64.5	39.5	44.8	70.4	39.85
LongLLaVA [52]	28.3	37.1	27.2	24.5	60.0	56.7	54.9	48.1	32.6	61.5	38.3	41.9	26.3	24.1	66.7	38.45
Kangaroo [30]	29.2	42.0	24.3	30.6	56.7	66.7	57.7	31.5	26.8	67.7	47.8	61.3	21.1	6.9	55.6	37.10
InternVL2-2B [6]	23.7	24.4	24.8	36.7	76.7	63.3	53.5	48.9	21.8	80.0	40.2	29.0	47.4	6.9	83.3	36.75
LongVILA [28]	25.1	35.6	40.6	40.8	80.0	60.0	38.0	32.8	25.1	76.9	20.6	64.5	50.0	37.9	79.6	36.46
Qwen2-VL-2B [49]	21.0	29.3	25.2	30.6	63.3	70.0	42.3	50.6	23.8	67.7	37.3	74.2	34.2	24.1	63.0	36.16
Video-LLaMA2-7B [8]	25.1	29.3	23.3	30.6	70.0	66.7	40.8	31.5	26.4	72.3	40.2	29.0	44.7	27.6	66.7	34.35
VILA-1.5-3B [28]	20.1	32.7	38.1	51.0	53.3	46.7	31.0	26.4	10.5	72.3	35.4	32.3	36.8	10.3	83.3	31.95
Video-LLaVA [27]	26.5	25.9	38.1	32.7	53.3	40.0	25.4	26.8	23.0	55.4	30.1	38.7	21.1	51.7	51.9	31.07
Chat-UniVi [19]	25.1	31.7	30.2	30.6	53.3	30.0	22.5	26.4	24.3	29.2	21.1	32.3	34.2	44.8	40.7	28.02
InternVL2-1B [6]	24.7	24.9	22.8	22.4	46.7	33.3	22.5	26.4	26.8	30.8	30.6	22.6	23.7	34.5	29.6	26.61
RANDOM	26.0	25.8	25.3	24.3	23.7	23.7	24.8	25.0	25.3	27.4	24.4	20.3	24.7	25.2	28.7	25.33

to the total questions across all sub-tasks:

$$\text{Overall ACC} = \frac{1}{\sum_i N_i} \sum_i \sum_{j=1}^{N_i} s_{q_j}. \quad (2)$$

3. Main Results

In this section, we analyze and quantify the video composition understanding capabilities of state-of-the-art MLLMs, providing a comprehensive evaluation of these models. For all experiments, we use a standardized prompt template and the default hyperparameters specified for each model.

Overall Performance. As shown in Table 2, the overall performance on the VIDCOMPOSITION benchmark reveals that understanding intricate video compositions remains challenging for MLLMs. While humans achieve exceptionally high scores (86.26), the leading models, such as LLaVA-OneVision-72B [23] (63.31), InternVL2-40B [6] (60.73), InternVL2-76B [6] (58.7) and Qwen2-VL-72B [49] (58.78), demonstrate only moderate success, underscoring the complexity of the tasks and the current limitations in video composition understanding. This gap between human and model performance highlights the benchmark's rigor and the need for advancements in fine-grained video-based compositional learning. Open-source models with advanced vision components, particularly InternVL2 variants, outperform API-based models like GPT-4o [1] (52.93) and Genmini-1.5-Flash [43] (52.40). The mean overall accuracy of these MLLMs is 43.44. For reference, the random-choice baseline has a score of 25.33. While the models exceed it, they still face considerable obstacles in approaching human-level video composition understanding.

Strengths & Weaknesses Analysis. From Table 2, we see that MLLMs generally perform better in **CU** tasks, particularly **A-P** and **CMP-P**. For example, top models like LLaVA-OneVision-72B [23], GPT-4o [1] and GPT-4o mini [1] achieve high scores in **CMP-P**; LLaVA-OneVision-72B [23] and InternVL2-40B [6] get 90.0 on **A-P**. This indicates that state-of-the-art MLLMs can effectively recognize and interpret actions and visual details of characters in a scene. Models also show strong performance on **SP** tasks such as **B-P** and **L-P**, with models like Gemini-1.5-Flash [43] achieves 83.1 on **B-P** and LLaVA-OneVision-72B [23] gets 90.3 on **L-P**. Additionally, these models achieve competitive scores on some **MA** tasks, such as **AS-P** and **SE-P**, with top scores reaching 89.5 and 85.2, respectively. This strong performance may be attributed to the fact that these tasks

Table 3. Resolution Analysis. Models are compared based on **#frm**, **LLM size**, and **Res.**. **CA**, **CU**, **NU**, **SP**, **MA**, and **Overall** are averaged on models in each row. The results indicate that higher **Res.** leads to improved performance in most cases, with highlighted relative gains.

Models	#frm	LLM size	Res.	CA	CU	NU	SP	MA	Overall
Chat-UniVi [19]; Video-LLaVA [27]; VideoChat2 [25]	8	7B	224	31.68	42.22	31.02	40.11	42.98	34.94
Chat-UniVi-v1.5 [19]; LongVA [61]; Video-LLaMA2 [8]			336	33.6$_{+1.92}$	48.89$_{+6.67}$	32.42$_{+1.4}$	**44.26**$_{+4.15}$	50.96$_{+7.98}$	38.04$_{+3.1}$
Video-LLaMA2.1 [8]; Video-LLaMA2.1-AV [8]			384	**36.9**$_{+3.3}$	**55.28**$_{+6.39}$	**46.0**$_{+13.58}$	43.11$_{-1.15}$	**53.72**$_{+2.76}$	**43.7**$_{+5.66}$
Chat-UniVi [19]; Video-LLaVA [27]; VideoChat2 [25]	16	7B	224	29.66	39.81	31.86	28.52	34.71	31.52
Chat-UniVi-v1.5 [19]; LongVA [61]; Video-LLaMA2 [8]			336	32.75$_{+3.09}$	49.07$_{+9.26}$	32.21$_{+0.35}$	**44.92**$_{+16.4}$	49.04$_{+14.33}$	37.67$_{+6.15}$
Video-LLaMA2.1 [8]; Video-LLaMA2.1-AV [8]			384	**37.06**$_{+4.31}$	**57.22**$_{+8.15}$	**45.68**$_{+13.47}$	43.44$_{-1.48}$	**54.13**$_{+5.09}$	**43.96**$_{+6.29}$
LongVA [61]; Video-LLaMA2 [8]	32	7B	336	31.39	46.67	35.26	**44.26**	53.72	37.98
Video-LLaMA2.1 [8]; Video-LLaMA2.1-AV [8]			384	**38.5**$_{+7.11}$	**59.72**$_{+13.05}$	**45.47**$_{+10.21}$	42.95$_{-1.31}$	**54.55**$_{+0.83}$	**44.64**$_{+6.66}$

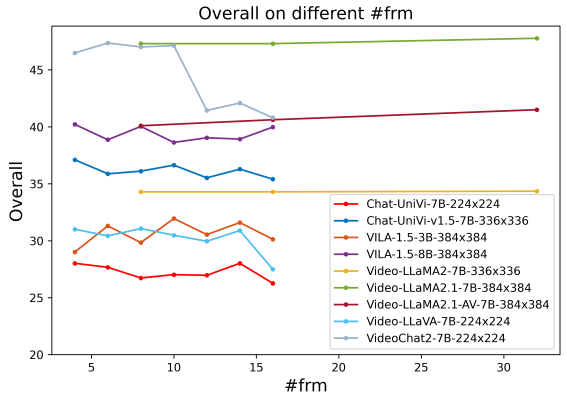

Figure 4. #frm analysis. We compare the overall accuracy of models from the same series with the same **LLM size** and **Res.**. The results indicate a counterintuitive irrelevance between the overall accuracy and input #frm.

rely on some expert knowledge about video-making techniques, which MLLMs can acquire from massive corpora. **Conversely, the models encounter significant difficulties in more complex compositional tasks, especially CA.** For example, **CamM-P** and **SS-P** yield only modest scores, with top models reaching 57.1 and 63.9, respectively, reflecting a significant gap in understanding cinematic techniques. **NU** tasks, such as **S-M** and **P-O**, also present challenges, with model performance substantially trailing behind human benchmarks because there is often a gap between scripts and actual video presentation. Unlike humans, who can intuitively bridge this gap, MLLMs are fine-tuned on closely matched vision-text pairs, limiting their ability to interpret subtle or implied connections in narrative tasks. Additionally, counting tasks, such as **Cha-C**, **S-C**, and **Cut-C**, remain particularly problematic for most models, further underscoring the limitations in visual counting abilities and understanding scene transitions across multiple frames across all models.

4. Diagnostic Analysis of Factors Affecting MLLMs' Performance

In this section, we analyze the factors that may affect the MLLM's understanding of video composition. We focus on four factors: the number of input frames (**#frm**), the resolution of the visual encoder (**Res.**), the size of the language decoder (**LLM size**), and the volume of training data (**Data volume**) in the SFT stage. We provide full analysis tables and figures of each factor in Supplementary.

Table 4. **Overall** accuracy of LongVA [61] and LongVILA [28] on different #frm ranging from 4 to 128.

Model	Res.	LLM size	#frm					
			4	8	16	32	64	128
LongVA [61]	336	7B	40.74	**43.73**	43.32	41.62	39.98	39.39
LongVILA [28]	Dynamic	8B	33.00	35.87	35.11	36.46	**36.17**	36.11

The Number of Input Frames. Across all the models, we consistently observe that the input frames don't contribute to the performance. As shown in the Figure 4 and Table 4, the overall accuracy is either stable or fluctuates randomly. We couldn't see any clear trends, although intuitively, extra frames would provide more information to help the model make decisions. We suspect that while more frames provide more information, this small amount of useful information is mixed in with a large amount of duplicate information, and the model cannot effectively extract it. This is against our expectation that it would bring benefits to counting tasks (**Cha-C**, **S-C**, **Cut-C**).

The Resolution of Visual Encoder. We observe that MLLMs with higher-resolution visual encoders perform significantly better. While the resolution is unchangeable for one specific model, we calculate the mean performance of all models with the same LLM size and video frames. As shown in Table 3, as resolution increases, performance on all five main categories increases consistently. However, it is worth noting that it is impossible to determine how much of this improvement is due to the higher-resolution visual encoder and how much is due to the different models themselves.

The Size of Language Decoder. To analyze this relationship more accurately, we compare models with different decoder sizes while keeping the encoder and training data constant. From Table 5, we observe that models with larger decoders demonstrate stronger performance, and the gains are mainly from **NU**, which requires the model not only to recognize individual frames but also to establish logical and causal

Table 5. LLM size analysis. We compare models from the same series with the same **#frm** and **Res.** but different **LLM sizes**. The results indicate that larger **LLM sizes** lead to improved performance in most cases, with highlighted relative gains.

Model	Res.	#frm	LLM size	CA	CU	NU	SP	MA	Overall
Qwen2-VL [49]	Dynamic	2 fps	2B	25.08	47.22	37.05	47.54	44.63	36.17
			7B	34.35$_{+9.27}$	56.67$_{+9.45}$	61.05$_{+24.0}$	57.38$_{+9.84}$	**48.76**$_{+4.13}$	49.3$_{+13.13}$
			72B	**50.48**$_{+16.13}$	**60.0**$_{+3.33}$	**71.16**$_{+10.11}$	**60.0**$_{+2.62}$	46.28$_{-2.48}$	**58.68**$_{+9.38}$
VILA-1.5 [28]	384	8	3B	26.84	43.89	19.16	37.38	47.11	29.84
			8B	**35.94**$_{+9.1}$	**52.78**$_{+8.89}$	**34.74**$_{+15.58}$	**42.62**$_{+5.24}$	**56.2**$_{+9.09}$	**40.04**$_{+10.2}$
		16	3B	26.04	41.67	23.37	34.75	48.76	30.13
			8B	**35.78**$_{+9.74}$	**51.11**$_{+9.44}$	**36.42**$_{+13.05}$	**39.34**$_{+4.59}$	**60.33**$_{+11.57}$	**39.98**$_{+9.85}$
Video-LLaMA2 [8]	336	32	7B	25.88	47.78	28.84	45.9	50.41	34.35
			72B	**54.15**$_{+28.27}$	**71.67**$_{+23.89}$	**65.68**$_{+36.84}$	**48.52**$_{+2.62}$	**59.5**$_{+9.09}$	**58.62**$_{+24.27}$
InternVL2 [7]	448	16	0.5B	24.12	28.33	26.53	29.84	28.93	26.61
			1.8B	24.28$_{+0.16}$	54.44$_{+26.11}$	35.16$_{+8.63}$	47.54$_{+17.7}$	53.72$_{+24.79}$	36.75$_{+10.14}$
			3.8B	32.11$_{+7.83}$	51.67$_{-2.77}$	44.0$_{+8.84}$	48.85$_{+1.31}$	48.76$_{-4.96}$	41.68$_{+4.93}$
			8B	**57.03**$_{+24.92}$	62.78$_{+11.11}$	53.68$_{+9.68}$	45.57$_{-3.28}$	56.2$_{+7.44}$	54.63$_{+12.95}$
			20B	40.1$_{-16.93}$	63.89$_{+1.11}$	51.16$_{-2.52}$	38.36$_{-7.21}$	54.55$_{-1.65}$	46.42$_{-8.21}$
			34B	54.95$_{+14.85}$	69.44$_{+5.55}$	**65.47**$_{+14.31}$	**56.07**$_{+17.71}$	**70.25**$_{+15.7}$	**60.73**$_{+14.31}$
			70B	51.92$_{-3.03}$	**72.78**$_{+3.34}$	64.84$_{-0.63}$	52.79$_{-3.28}$	63.64$_{-6.61}$	58.73$_{-2.0}$

Table 6. Data volume analysis. We compare models with the same **#frm**, **Res.**, and **LLM sizes** but using different **Data volume** in the SFT stage. The results indicate that larger **Data volumes** lead to improved performance in most cases, with highlighted relative gains.

Model	#frm	Res.	LLM size	Data volume	CA	CU	NU	SP	MA	Overall
Chat-UniVi [19]	8	224	7B	0.65M	25.56	34.44	23.37	25.9	36.36	26.73
VideoChat2 [25]				2M	**39.46**$_{+13.9}$	**57.78**$_{+23.34}$	**44.84**$_{+21.47}$	**58.03**$_{+32.13}$	**50.41**$_{+14.05}$	**47.01**$_{+20.28}$
Chat-UniVi-v1.5 [19]	8	336	7B	1.27M	37.38	47.78	26.53	37.38	46.28	36.11
LongVA [61]				1.32M	**37.54**$_{+0.16}$	**51.67**$_{+3.89}$	**41.89**$_{+15.36}$	**49.51**$_{+12.13}$	**56.2**$_{+9.92}$	**43.73**$_{+7.62}$
VILA-1.5 [28]	8	384	8B	1.21M	35.94	52.78	34.74	42.62	56.2	40.04
Video-LLaMA2.1-AV [8]			7B	3.35M	35.46$_{-0.48}$	52.78$_{+0}$	37.05$_{+3.31}$	39.34$_{-3.28}$	**58.68**$_{+2.48}$	40.09$_{+0.05}$
Video-LLaMA2.1 [8]			7B	3.35M	**38.34**$_{+2.88}$	**57.78**$_{+5.0}$	**54.95**$_{+17.9}$	**46.89**$_{+7.75}$	48.76$_{-9.92}$	**47.3**$_{+7.22}$
Chat-UniVi [19]	16	224	7B	0.65M	24.92	32.22	23.37	24.92	**38.84**	26.26
VideoChat2 [25]				2M	**39.3**$_{+14.38}$	**60.0**$_{+27.78}$	**44.84**$_{+21.47}$	**31.8**$_{+6.88}$	26.45$_{-12.39}$	**40.8**$_{+14.54}$
Chat-UniVi-v1.5 [19]	16	336	7B	1.27M	34.5	48.89	25.89	40.98	42.98	35.4
LongVA [61]				1.32M	**37.86**$_{+3.36}$	**51.11**$_{+2.22}$	**41.89**$_{+16.0}$	**47.87**$_{+6.89}$	**53.72**$_{+10.74}$	**43.32**$_{+7.92}$
VILA-1.5 [28]	16	384	8B	1.21M	35.78	51.11	36.42	39.34	**60.33**	39.98
Video-LLaMA2.1-AV [8]			7B	3.35M	35.78$_{+0}$	56.67$_{+5.56}$	36.42$_{+0}$	40.0$_{+0.66}$	59.5$_{-0.83}$	40.62$_{+0.64}$
Video-LLaMA2.1 [8]			7B	3.35M	**38.34**$_{+2.56}$	**57.78**$_{+1.11}$	**54.95**$_{+18.53}$	**46.89**$_{+6.89}$	48.76$_{-10.74}$	**47.3**$_{+6.68}$
Kangaroo [30]	64	448	8B	2.94M	31.79	51.67	29.05	**53.44**	33.06	37.1
MiniCPM-V [55]				8.32M	**38.18**$_{+6.39}$	**60.0**$_{+8.33}$	**40.84**$_{+11.79}$	38.36$_{-15.08}$	**55.37**$_{+22.31}$	**42.5**$_{+5.4}$

relationships across sequences, a capability that benefits from a more powerful language decoder. Tasks in **MA** also benefit from the external knowledge acquired by LLMs.

The Volume of Training Data. We can observe the performance influence brought by fine-tuning the MLLMs with more data from Table 6. We compare models with the same configuration, *e.g.* the same number of input frames, the same resolution of the vision encoder, and the same or similar size of LLM adopted. The results indicate that larger data volumes lead to improved performance of video composition understanding in most cases.

Qualitative Analysis. We perform an error analysis to gain deeper insights into the models' shortcomings in fine-grained video composition understanding. In this analysis, the models are required to answer questions and provide explanations in a dialogue format. Figure 5 shows the examples where top models fail to predict correct answers. For example, while humans easily use visual context to distinguish camera movements and angles like "pan left" and "zoom in" or angles like "eye level" and "low angle," models like LLaVA-OneVision-72B [23] and GPT-4 [1] often struggle due to scene transitions and subtle perspective changes.

5. Related Work

MLLMs for Video Understanding. Equipping LLMs with adapted video encoders has led to the creation of several multimodal models tailored for video understanding [46]. For instance, GPT-4-turbo, GPT-4o, and GPT-4o-mini [1] are GPT-based models with integrated video comprehension capabilities. InternVL2 [7], with parameter counts ranging from 1B to 76B, is based on the InternLM framework [4] and supports video processing at multiple scales. Additionally, models derived from the LLaMA backbone—such as LLaVA-OneVision [23], VILA [28], VideoLLaMA [8], and LongLLaVA [52]—have been adapted for video input.

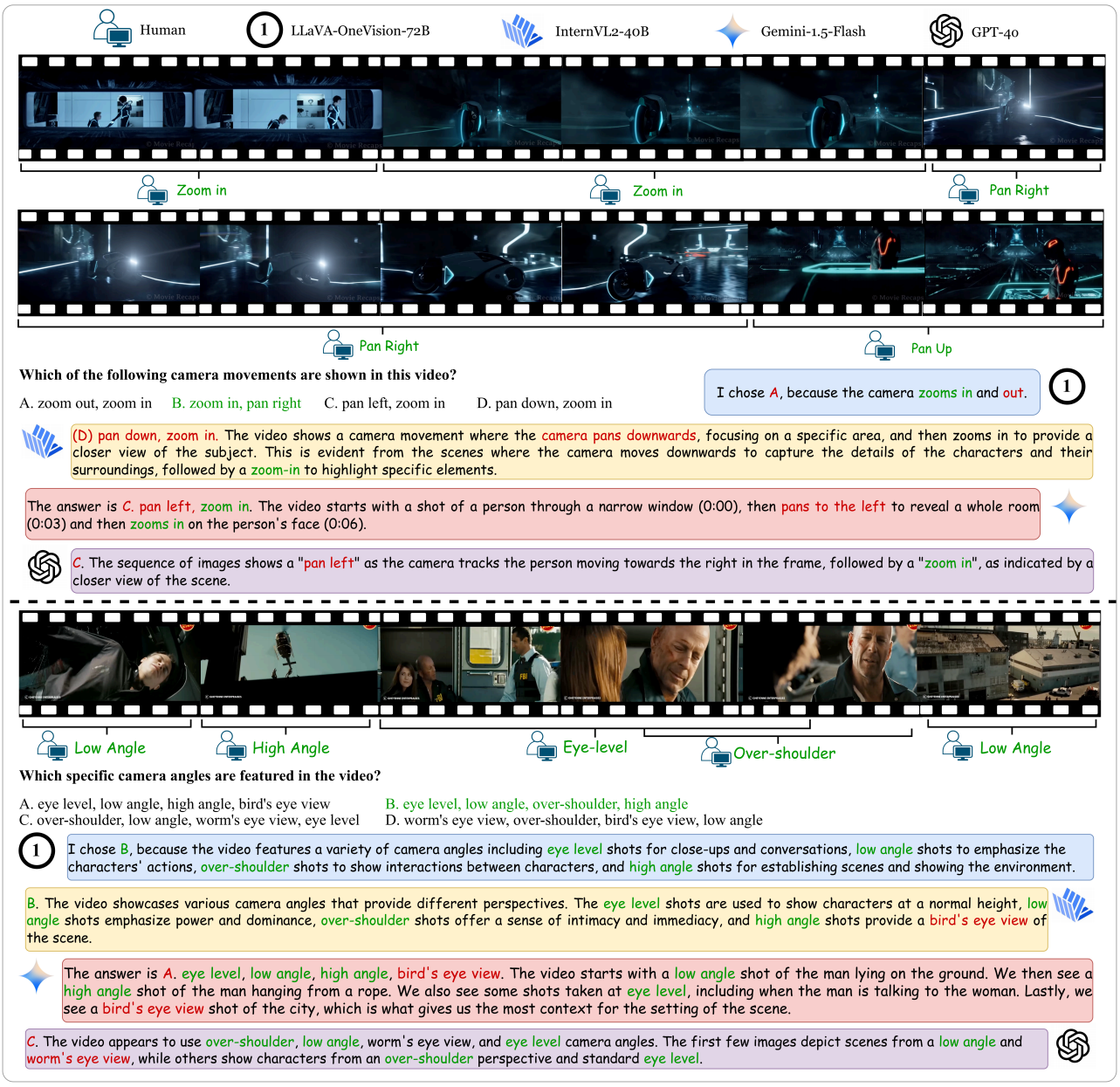

Figure 5. Qualitative analysis. Green represents correct answers, while red indicates wrong prediction or explanation. More cases can be found in Supplementary.

Gemini has also been extended to include video processing capabilities [43]. Other models, including Qwen [49], MiniCPM [55], Kangaroo [30], and Chat-UniVi [19], exhibit strong video understanding abilities. In our work, we thoroughly evaluate these models' capabilities in video composition understanding and provide detailed analysis.

Evaluation Benchmarks for MLLMs. Numerous benchmarks for MLLMs have recently emerged to evaluate diverse model capabilities, with image captioning, Visual Question Answering (VQA), and visual reasoning among the most frequently assessed tasks. Image captioning [9, 29, 38, 42] measures an MLLM's ability to generate text descriptions of visual content. VQA [2, 37, 39] assesses the model's proficiency in answering questions based on visual inputs by integrating visual perception with language understanding and external knowledge. Visual reasoning [17, 20, 45] evaluates a model's spatial awareness and logical reasoning in processing visual information. Moreover, the comprehensive abilities of MLLMs are gauged using advanced benchmarks [11, 13, 22, 31, 35, 36, 56, 57, 59]. For video MLLMs, similar efforts leverage existing benchmarks [26, 51] to evaluate video understanding [24, 41]. However, a notable gap remains in evaluating models on the video-composition under-

standing. This composition understanding is crucial for accurately processing and correlating multiple elements within a visual scene [14, 60]. While existing benchmarks assess compositionally in images [16, 48], few comprehensively address the specific challenges of compositional understanding in video, where many MLLMs still show limitations.

6. Conclusion

We introduce VIDCOMPOSITION, a novel and high-quality benchmark designed to evaluate MLLMs in understanding video compositions. Our benchmark incorporates various video types and QA categories, covering various aspects of video composition, *e.g.* camera movement, shot size, narrative structure, and character actions. Through VIDCOMPOSITION, we comprehensively assess MLLMs' abilities to understand complex video compositions. The evaluation reveals a significant performance gap between humans and models, shedding light on the limitations of current MLLMs and providing valuable insights for future improvements.

Acknowledgement

This work was supported in part by the National Eye Institute of the National Institutes of Health under award number R01EY034562 and the Defense Advance Research Projects Agency under contract number HR00112220003. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies; no official endorsement should be inferred.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 5, 7

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 8

[3] Jing Bi, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. Eagle: Egocentric aggregated language-video engine. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 1682–1691, New York, NY, USA, 2024. Association for Computing Machinery. 1

[4] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 7

[5] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. *arXiv preprint arXiv:2311.14906*, 2023. 2, 3

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 5

[7] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 7

[8] 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. 1, 3, 5, 6, 7

[9] Mingqian Feng, Yunlong Tang, Zeliang Zhang, and Chenliang Xu. Do more details always introduce more hallucinations in lvlm-based image captioning? *arXiv preprint arXiv:2406.12663*, 2024. 8

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 2, 3

[11] Chaoyou Fu, Peixian Chen, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 8

[12] 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. 1, 2, 3

[13] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 8

[14] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 36, 2024. 9

[15] Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*, 2024. 1

[16] Hang Hua, Yunlong Tang, Ziyun Zeng, Liangliang Cao, Zhengyuan Yang, Hangfeng He, Chenliang Xu, and Jiebo Luo. Mmcomposition: Revisiting the compositionality of pre-trained vision-language models. *arXiv preprint arXiv:2410.09733*, 2024. 2, 3, 9

[17] Hang Hua, Jing Shi, Kushal Kafle, Simon Jenni, Daoan Zhang, John Collomosse, Scott Cohen, and Jiebo Luo. Finematch: Aspect-based fine-grained image and text mismatch detection and correction. In *European Conference on Computer Vision*, pages 474–491. Springer, 2025. 8

[18] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 3

[19] 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. 5, 6, 7, 8

[20] 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*, 2017. 8

[21] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 2, 3

[22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 8

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 3, 5, 7

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. 8

[25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 2, 3, 5, 6, 7

[26] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, Tamara Lee Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. Value: A multi-task benchmark for video-and-language understanding evaluation, 2021. 8

[27] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 5, 6

[28] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 3, 5, 6, 7

[29] 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 *ECCV*. Springer, 2024. 8

[30] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024. 5, 7, 8

[31] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2, 8

[32] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 2, 3

[33] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 1, 3

[34] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution, 2024. 5

[35] Jian Lu, Shikhar Srivastava, Junyu Chen, Robik Shrestha, Manoj Acharya, Kushal Kafle, and Christopher Kanan. Revisiting multi-modal llm evaluation. *arXiv preprint arXiv:2408.05334*, 2024. 8

[36] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 8

[37] 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. 8

[38] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv*, abs/2203.10244, 2022. 8

[39] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2020. 8

[40] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 2, 3

[41] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023. 8

[42] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *arXiv:2404.19753*, 2024. 8

[43] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Alayrac, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 3, 5, 8

[44] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 2, 3

[45] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2017. 8

[46] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023. 7

[47] Yunlong Tang, Daiki Shimada, Jing Bi, and Chenliang Xu. Avicuna: Audio-visual llm with interleaver and context-boundary alignment for temporal referential dialogue. *arXiv preprint arXiv:2403.16276*, 2024. 1

[48] 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. 2, 3, 9

[49] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3, 5, 7, 8

[50] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024. 2, 3

[51] 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, 2020. 8

[52] Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. Longllava: Scaling multi-modal llms to 1000 images efficiently via hybrid architecture. *arXiv preprint arXiv:2409.02889*, 2024. 5, 7

[53] 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 (CVPR)*, pages 9777–9786, 2021. 2, 3

[54] 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. 2, 3

[55] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5, 7, 8

[56] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024. 8

[57] Yongsheng Yu, Ziyun Zeng, Hang Hua, Jianlong Fu, and Jiebo Luo. Promptfix: You prompt and we fix the photo. *arXiv preprint arXiv:2405.16785*, 2024. 8

[58] 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*, pages 9127–9134, 2019. 2, 3

[59] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023. 8

[60] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022. 9

[61] 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. 5, 6, 7