VIDCOMPOSITION: Can MLLMs Analyze Compositions in Compiled Videos?

Supplementary Material

8. More Statistics for VIDCOMPOSITION

In this section, we show more statistics for VIDCOMPOSITION. Figure 6 presents a word cloud that highlights key

Figure 6. Word cloud highlighting key terms from questions and options, showcasing the diversity of video compositions included in our benchmark.

terms from the questions and options, demonstrating the diversity of video compositions included in our benchmark. Figure 7 illustrates the distribution of video frames across

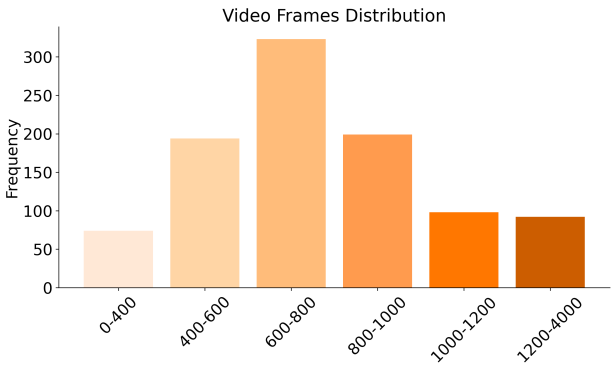

Figure 7. Distribution of video frames across different ranges, illustrating the varying durations of videos in our benchmark.

different ranges, highlighting the diversity in video durations present in our dataset. Most videos are concentrated in the 600-800 frame range, with fewer videos having shorter or longer durations. This distribution reflects a balanced yet diverse set of videos, suitable for comprehensive benchmarking.

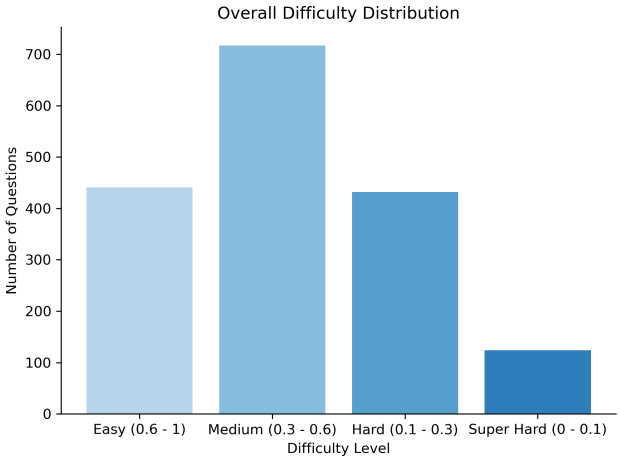

Figure 8. Distribution of questions across different difficulty levels, ranging from "Easy" (answered correctly by >60% of models,) to "Super Hard" (answered correctly by <10% of models).

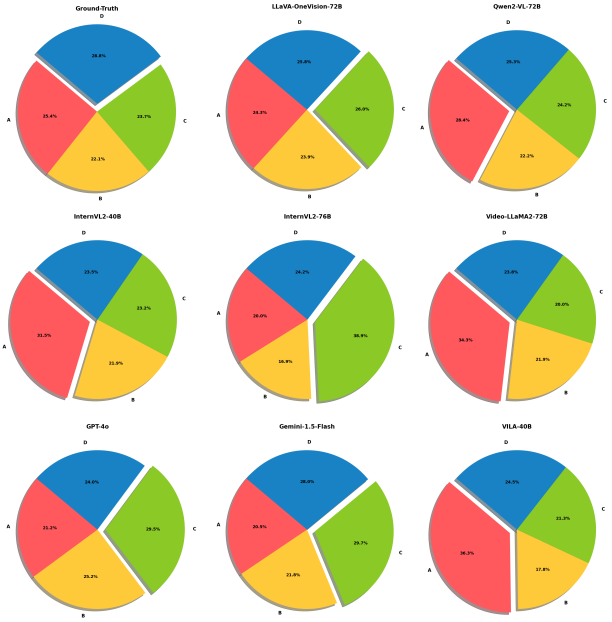

Figure 9. The distribution of answers for human-annotated questions is relatively balanced, as shown in the Ground-Truth pie chart. Predictions from several top models also exhibit relatively balanced distributions.

Figure 8 presents the distribution of questions across four difficulty levels: "Easy" (answered correctly by more than 60% of models), "Medium" (answered correctly by 30%-60% of models), "Hard" (answered correctly by 10%-30% of

models), and "Super Hard" (answered correctly by less than 10% of models). The pie charts in Figure 9 show that the answers for human-annotated questions are distributed fairly evenly across all options. Similarly, predictions from several top-performing models also demonstrate a comparable level of balance in their distributions. Comparing with the main results table, it can be observed that better model performance correlates with more evenly distributed predictions. For example, InternVL-40B outperforms InternVL-76B, as the pie chart reveals that InternVL-76B predictions are less evenly distributed and are biased toward option C compared to InternVL-40B. As shown in Figure 14, the tendency for imbalanced predictions is more pronounced in models with smaller sizes of LLM.

Figure 10 provides a detailed view of the distribution of questions across four difficulty levels (Easy, Medium, Hard, Super Hard) for each sub-task, reflecting the diverse challenges within the benchmark.

9. Tasks Definition in VIDCOMPOSITION

- **Camera Movement Perception (CamM-P)**: Identifying the types of camera movements shown in a video, such as panning, zooming, or tracking, which affect the visual flow and dynamics of the scene.
- **Shot Size Perception (SS-P)**: Recognizing the shot sizes, like close-up, medium shot, or full shot, which contribute to the viewer's sense of intimacy or scope within the scene.
- **Camera Angle Perception (CamA-P)**: Identifying different camera angles used in the video, such as low angle, bird's-eye view, or over-shoulder, which impact the perspective and interpretive context of the scene.
- **Emotion Perception (E-P)**: Detecting the emotions displayed by characters, such as fear, sadness, or happiness, which contribute to narrative understanding and character development.
- **Action Perception (A-P)**: Recognizing actions performed by characters, like driving or talking, to understand the physical activities and plot progression in the video.
- **Costume, Makeup, and Props Perception (CMP-P)**: Identifying elements of costume, makeup, and props used by characters, which provide contextual and stylistic cues about the setting, era, or genre.
- **Character Counting (Cha-C)**: Counting the number of characters appearing in the video, which gives an understanding of scene complexity and interaction density.
- **Script Matching (S-M)**: Identifying the narrative script or dialogue that corresponds with the visual content, facilitating alignment of visual and textual story elements.
- **Plot Ordering (P-O)**: Determining the chronological sequence of events depicted in the video, enabling coherent understanding of the storyline and causality.
- **Background Perception (B-P)**: Recognizing the type of background setting, such as a lakeside or grassland, which anchors the scene's location and environmental context.
- **Scene Counting (S-C)**: Counting the distinct scenes or settings within the video, indicating shifts in location or time that structure the narrative.
- **Lighting Perception (L-P)**: Identifying lighting conditions in the video, like high-key or low-key lighting, which affect the mood, visibility, and aesthetic of the scenes.
- **Art Style Perception (AS-P)**: Recognizing the art style of the video, such as Japanese cel anime or 3D CG animation, which contributes to the visual genre and artistic tone.
- **Cut Counting (Cut-C)**: Counting the number of cuts in the video, which reflects editing style and pacing, impacting the rhythm and viewer engagement.
- **Special Effect Perception (SE-P)**: Identifying special effects used in the video, like explosions or rain, which add dramatic or fantastical elements to enhance the visual experience.

10. Prompt Template

The prompt template provides explicit instructions for selecting the correct answer from multiple-choice options based on the provided video. It is carefully crafted to minimize ambiguity and guide the model's reasoning process. By embedding a professional perspective (*e.g.*, "like a director and cinematographer"), the prompt attempts to align the model's decision-making with human-like attention to detail. Additionally, the rigid structure ensures the compatibility of the generated responses, with only "A," "B," "C," or "D" in the output.

Prompt Template for Model Prediction

Given a video, a multiple-choice question, and several options, ensure you select the option that correctly answers the question based on the provided video. Please consider comprehensively and meticulously like a professional director and cinematographer. Answer with the option's letter from the given choices directly, and don't contain any other contents!
{question}
{options}

In qualitative analysis, we also ask models to explain their answers. At this time, we replace "Answer with the option's letter from the given choices directly, and don't contain any other contents!" with "The output should contain the option index and explain why you selected this as your answer."

11. Annotation & Human Evaluation System

The annotation checker user interface is shown as Figure 11. The system is designed to ensure high-quality annotations in the benchmark. The user interface displays the video,

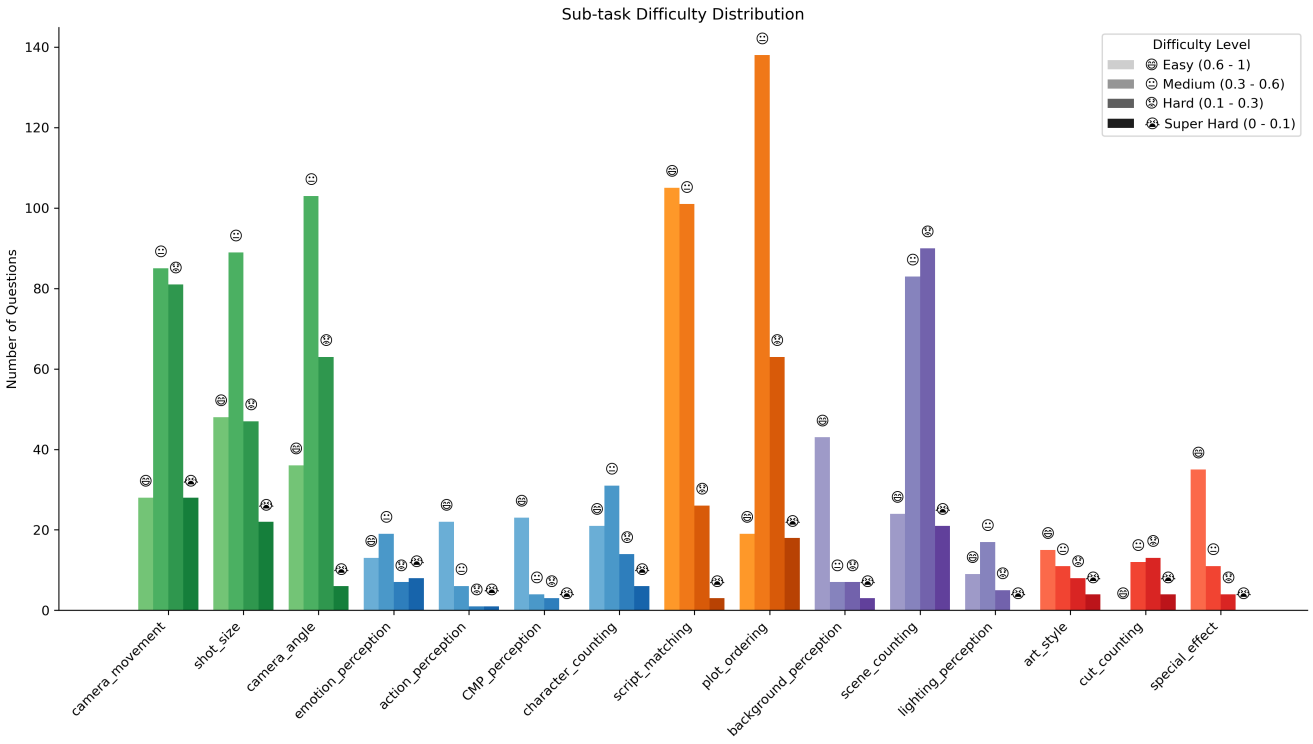

Figure 10. The difficulty distribution across various tasks in the benchmark. The bars represent the number of questions categorized into different difficulty levels (Easy, Medium, Hard, Super Hard) for each sub-task, highlighting variations in difficulty distribution across tasks.

the associated question, and multiple-choice options. It also includes a feedback section at the bottom, where reviewers can provide corrections or specify errors in the questions or options. Reviewers use this interface to assess the clarity and accuracy of annotations by attempting the questions themselves. If any issue is identified, such as unclear phrasing or incorrect answer options, they can use the feedback section to suggest improvements or note discrepancies. In addition to annotation refinement, this system is also used for human evaluation tasks, enabling consistent validation of both the dataset and the benchmark design. This iterative process ensures reliability, reduces errors, and supports the continual enhancement of question clarity and dataset quality.

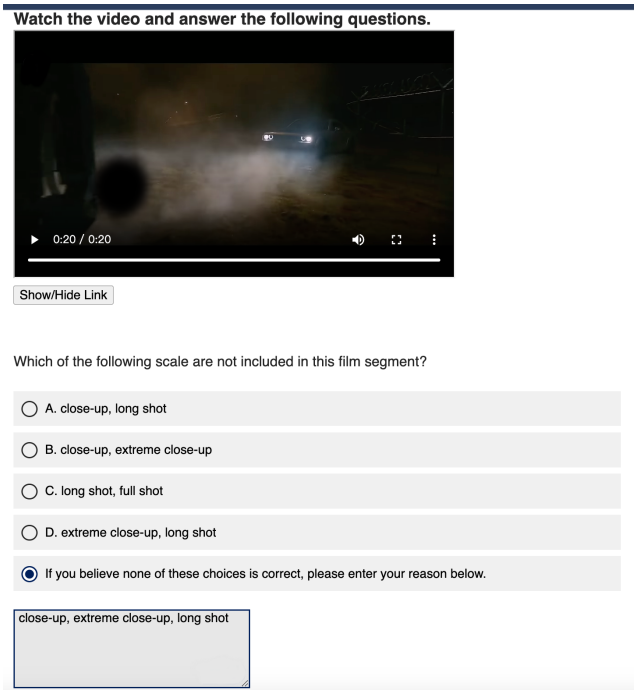

Figure 11. The user interface for annotation checker. It can also be used for human evaluation.

12. More Results

In this section, we present the complete results of the diagnostic analysis on factors influencing the performance of MLLMs on VIDCOMPOSITION. Specifically, the impact of the number of frames is illustrated in Figure 12, the resolution of the visual encoder is detailed in Table 7, the size of the LLM is analyzed in Table 8, and the effect of training data volume is shown in Table 9. We also provide more visualization results in this section.

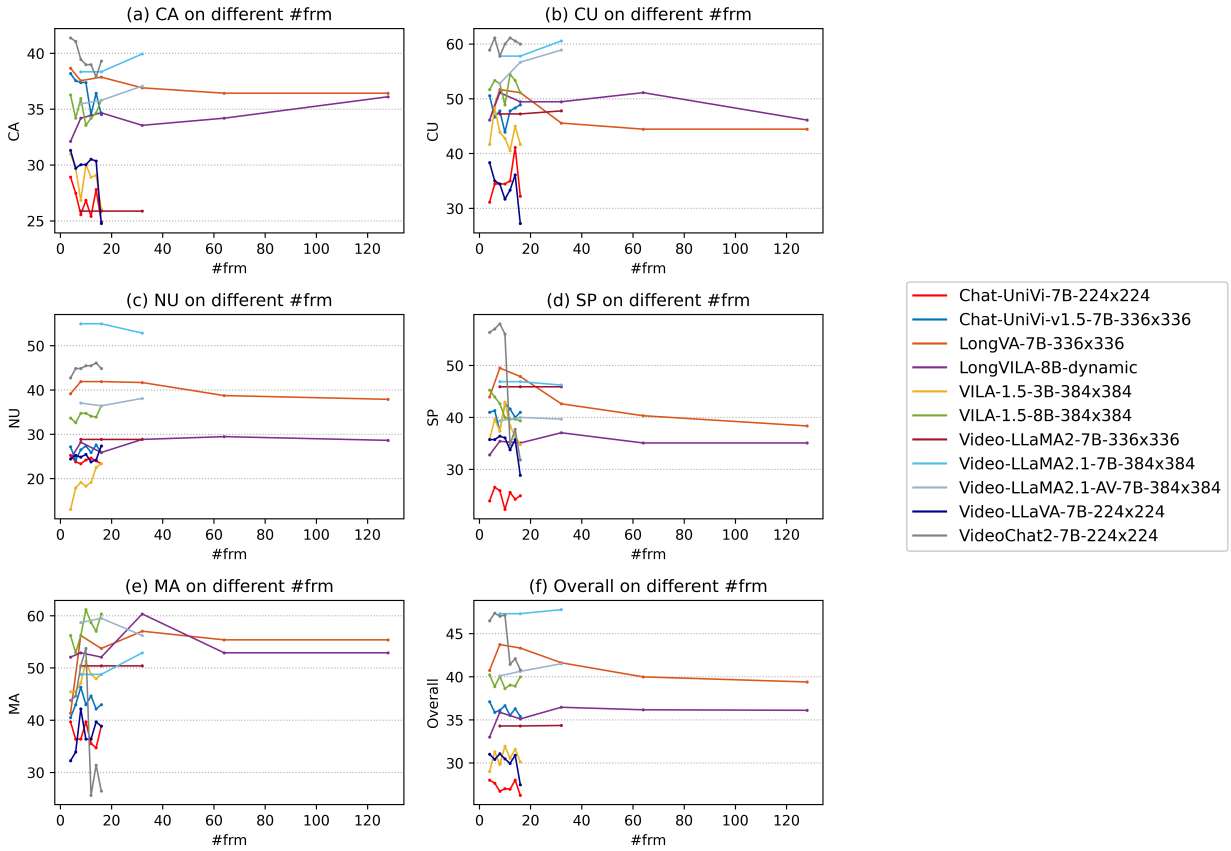

Figure 12. Full frame analysis. Performance analysis of models across different numbers of input frames. The results show no clear trends, with performance either remaining stable or fluctuating randomly as the number of frames increases.

Table 7. Full resolution analysis.

Models	#frm	LLM size	Res.	CA	CU	NU	SP	MA	Overall
Chat-UniVi; Video-LLaVA; VideoChat2	4	7B	224	33.87	42.78	30.81	38.69	38.57	35.17
Chat-UniVi-v1.5; LongVA			336	**38.42**$_{+4.55}$	**48.33**$_{+5.55}$	**33.16**$_{+2.35}$	**42.46**$_{+3.77}$	**40.91**$_{+2.34}$	**38.92**$_{+3.75}$
Chat-UniVi; Video-LLaVA; VideoChat2	6	7B	224	32.75	43.52	**31.3**	39.78	38.29	35.15
Chat-UniVi-v1.5			336	**37.54**$_{+4.79}$	**46.67**$_{+3.15}$	24.21$_{-7.09}$	**41.31**$_{+1.53}$	**42.98**$_{+4.69}$	**35.87**$_{+0.72}$
Chat-UniVi; Video-LLaVA; VideoChat2	8	7B	224	31.68	42.22	31.02	40.11	42.98	34.94
Chat-UniVi-v1.5; LongVA; Video-LLaMA2			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; Video-LLaMA2.1-AV			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; Video-LLaVA; VideoChat2	10	7B	224	31.95	42.04	**31.72**	38.14	**43.25**	34.88
Chat-UniVi-v1.5			336	**37.38**$_{+5.43}$	**43.89**$_{+1.85}$	27.37$_{-4.35}$	**42.62**$_{+4.48}$	42.98$_{-0.27}$	**36.64**$_{+1.76}$
Chat-UniVi; Video-LLaVA; VideoChat2	12	7B	224	31.63	43.15	**31.3**	31.37	32.51	32.79
Chat-UniVi-v1.5			336	**34.5**$_{+2.87}$	**47.78**$_{+4.63}$	25.89$_{-5.41}$	**41.64**$_{+10.27}$	**44.63**$_{+12.12}$	**35.52**$_{+2.73}$
Chat-UniVi; Video-LLaVA; VideoChat2	14	7B	224	32.0	45.93	**31.44**	32.57	35.26	33.67
Chat-UniVi-v1.5			336	**36.42**$_{+4.42}$	**48.33**$_{+2.4}$	27.58$_{-3.86}$	**40.0**$_{+7.43}$	**42.15**$_{+6.89}$	**36.28**$_{+2.61}$
Chat-UniVi; Video-LLaVA; VideoChat2	16	7B	224	29.66	39.81	31.86	28.52	34.71	31.52
Chat-UniVi-v1.5; LongVA; Video-LLaMA2			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; Video-LLaMA2.1-AV			384	**37.06**$_{+4.31}$	**57.22**$_{+8.15}$	**45.68**$_{+13.47}$	43.44$_{-1.48}$	**54.13**$_{+5.09}$	**43.96**$_{+6.29}$
VILA-1.5	16	8B	384	35.78	51.11	36.42	39.34	**60.33**	39.98
InternVL2			448	**57.03**$_{+21.25}$	**62.78**$_{+11.67}$	**53.68**$_{+17.26}$	**45.57**$_{+6.23}$	56.2$_{-4.13}$	**54.63**$_{+14.65}$
LongVA; Video-LLaMA2	32	7B	336	31.39	46.67	35.26	**44.26**	53.72	37.98
Video-LLaMA2.1; Video-LLaMA2.1-AV			384	**38.5**$_{+7.11}$	**59.72**$_{+13.05}$	**45.47**$_{+10.21}$	42.95$_{-1.31}$	**54.55**$_{+0.83}$	**44.64**$_{+6.66}$

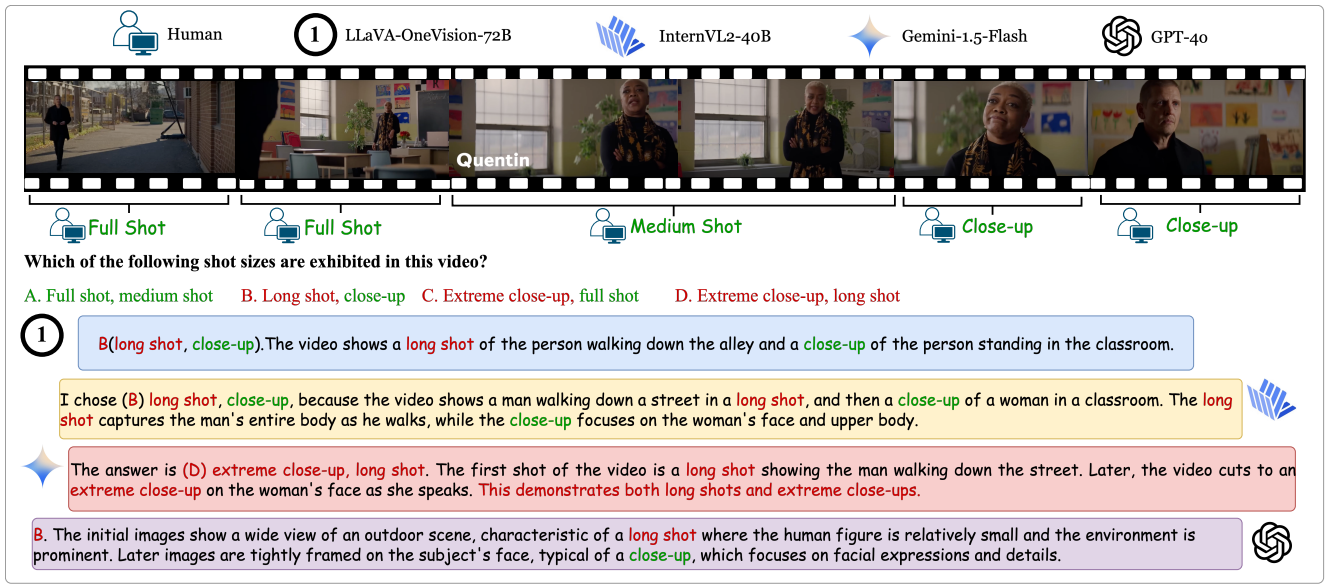

Figure 13. Example of a multiple-choice question assessing **SS-P** in a video. The figure compares predictions from human annotators and various models, highlighting their selected answers and reasoning processes. Correct answers and explanations are provided for reference.

Table 8. Full LLM size analysis.

Model	Res.	#frm	LLM size	CA	CU	NU	SP	MA	Overall
InternVL2	Dynamic	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$_{-2.0}$
Qwen2-VL	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	384	4	3B	30.99	41.67	13.05	35.74	45.45	29.02
			8B	**36.26**$_{+5.27}$	**51.67**$_{+10.0}$	**33.68**$_{+20.63}$	**45.25**$_{+9.51}$	**56.2**$_{+10.75}$	**40.21**$_{+11.19}$
		6	3B	29.71	48.33	17.89	39.67	45.45	31.3
			8B	**34.19**$_{+4.48}$	**53.33**$_{+5.0}$	**32.63**$_{+14.74}$	**43.93**$_{+4.26}$	**52.89**$_{+7.44}$	**38.86**$_{+7.56}$
		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**$_{+10.2}$
		10	3B	30.03	42.78	18.32	42.95	51.24	31.95
			8B	33.55$_{+3.52}$	48.89$_{+6.11}$	34.74$_{+16.42}$	40.0$_{-2.95}$	61.16$_{+9.92}$	38.63$_{+6.68}$
		12	3B	28.91	40.56	19.16	38.36	48.76	30.54
			8B	**34.19**$_{+5.28}$	**54.44**$_{+13.88}$	**34.11**$_{+14.95}$	**39.67**$_{+1.31}$	**58.68**$_{+9.92}$	**39.04**$_{+8.5}$
		14	3B	29.07	45.0	22.53	36.39	47.93	31.59
			8B	**34.66**$_{+5.59}$	**53.33**$_{+8.33}$	**33.89**$_{+11.36}$	**39.67**$_{+3.28}$	**57.02**$_{+9.09}$	**38.92**$_{+7.33}$
		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	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**$_{+24.27}$

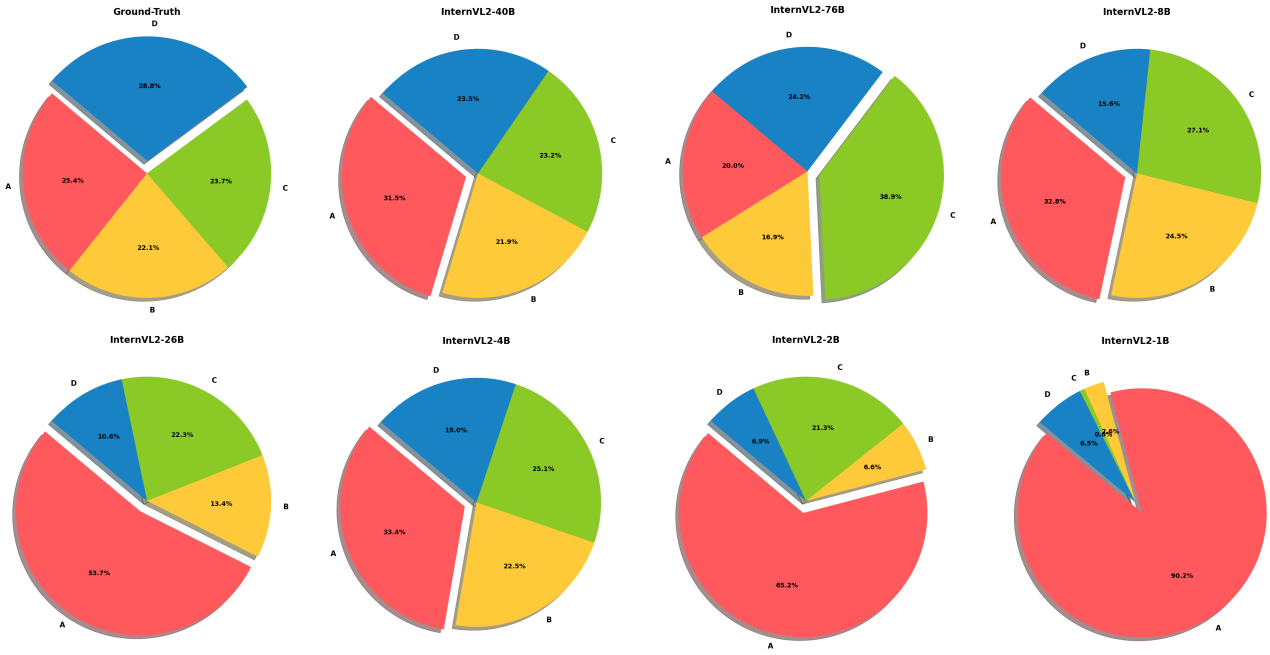

Figure 14. The pie charts show the answer distributions of InternVL models with varying LLM sizes. Larger models usually predict more balanced distributions, while smaller models, like InternVL-1B, exhibit strong biases, particularly toward option A.

Table 9. Full Data volume analysis.

Model	#frm	Res.	LLM size	Data volume	CA	CU	NU	SP	MA	Overall
Chat-UniVi	4	224	7B	0.65M	28.91	31.11	25.26	23.93	39.67	28.02
VideoChat2				2M	**41.37**$_{+12.46}$	**58.89**$_{+27.78}$	**42.74**$_{+17.48}$	**56.39**$_{+32.46}$	**43.8**$_{+4.13}$	**46.48**$_{+18.46}$
Chat-UniVi-v1.5	4	336	7B	1.27M	38.18	**50.56**	27.16	40.98	40.5	37.1
LongVA				1.32M	**38.66**$_{+0.48}$	46.11$_{-4.45}$	**39.16**$_{+12.0}$	**43.93**$_{+2.95}$	**41.32**$_{+0.82}$	**40.74**$_{+3.64}$
Chat-UniVi	6	224	7B	0.65M	27.48	34.44	23.79	26.56	36.36	27.67
VideoChat2				2M	**41.05**$_{+13.57}$	**61.11**$_{+26.67}$	**44.84**$_{+21.05}$	**57.05**$_{+30.49}$	**44.63**$_{+8.27}$	**47.36**$_{+19.69}$
Chat-UniVi	8	224	7B	0.65M	25.56	34.44	23.37	25.9	36.36	26.73
VideoChat2				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	8	336	7B	1.27M	37.38	47.78	26.53	37.38	46.28	36.11
LongVA				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	8	384	8B	1.21M	35.94	52.78	34.74	42.62	56.2	40.04
Video-LLaMA2.1-AV			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			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	10	224	7B	0.65M	26.84	34.44	24.21	22.3	39.67	27.02
VideoChat2				2M	**38.98**$_{+12.14}$	**60.0**$_{+25.56}$	**45.47**$_{+21.26}$	**56.07**$_{+33.77}$	**53.72**$_{+14.05}$	**47.13**$_{+20.11}$
Chat-UniVi	12	224	7B	0.65M	25.4	35.0	24.63	25.57	**35.54**	26.96
VideoChat2				2M	**38.98**$_{+13.58}$	**61.11**$_{+26.11}$	**45.47**$_{+20.84}$	**34.75**$_{+9.18}$	25.62$_{-9.92}$	**41.44**$_{+14.48}$
Chat-UniVi	14	224	7B	0.65M	27.8	41.11	24.0	24.26	**34.71**	28.02
VideoChat2				2M	**37.86**$_{+10.06}$	**60.56**$_{+19.45}$	**46.11**$_{+22.11}$	**37.7**$_{+13.44}$	31.4$_{-3.31}$	**42.09**$_{+14.07}$
Chat-UniVi	16	224	7B	0.65M	24.92	32.22	23.37	24.92	**38.84**	26.26
VideoChat2				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	16	336	7B	1.27M	34.5	48.89	25.89	40.98	42.98	35.4
LongVA				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	16	384	8B	1.21M	35.78	51.11	36.42	39.34	**60.33**	39.98
Video-LLaMA2.1-AV			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			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	64	448	8B	2.94M	31.79	51.67	29.05	**53.44**	33.06	37.1
MiniCPM-V				8.32M	**38.18**$_{+6.39}$	**60.0**$_{+8.33}$	**40.84**$_{+11.79}$	38.36$_{-15.08}$	**55.37**$_{+22.31}$	**42.5**$_{+5.4}$

Parse single letter choice

```python
def extract_single_content(text):
    # If text is a list, convert it to a string by taking the first element
    if isinstance(text, list):
        if text: # Ensure the list is not empty
            text = text[0]
        else:
            return random.choice(['A', 'B', 'C', 'D']) # Return default if list is empty

    # Check if text is a valid string
    if not isinstance(text, str):
        return random.choice(['A', 'B', 'C', 'D'])

    # 1. Match patterns like (A)(B)(C)(D)
    match = re.search(r'\((A|B|C|D)\)', text)
    if match:
        return match.group(1)

    # 2. Match text starting with A, B, C, or D, followed by spaces or non-alphabetic characters
    match = re.match(r'^(A|B|C|D)[\s\W]*', text)
    if match:
        return match.group(1)

    # 3. Match standalone A, B, C, or D
    match = re.match(r'\b[A-D]\b', text)
    if match:
        return match.group(1)

    # 4. Match patterns like (a), (b), (c), (d) and convert to uppercase
    match = re.search(r'\((a|b|c|d)\)', text)
    if match:
        return match.group(1).upper()

    # 5. Match patterns like A., B., C., or D.
    match = re.search(r'\b(A|B|C|D)\.', text)
    if match:
        return match.group(1)

    # 6. If text contains a single letter, return it in uppercase
    letters = re.findall(r'[a-zA-Z]', text)
    if len(letters) == 1:
        return letters[0].upper()

    # Default return if no patterns match
    return random.choice(['A', 'B', 'C', 'D'])
```