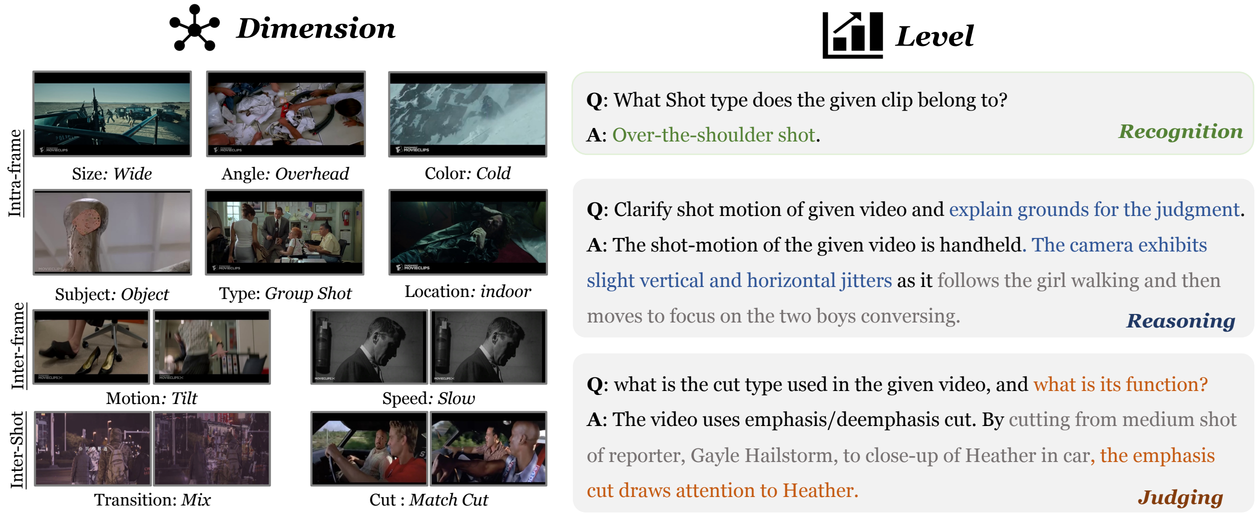
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VEU-Bench: Towards Comprehensive Understanding of Video Editing

Bozheng Li[1,2] Yongliang Wu[1,3] Yi Lu[1,4] Jiashuo Yu[5] Licheng Tang[1] Jiawang Cao[1]
Wenqing Zhu[1] Yuyang Sun[1] Jay Wu[1] Wenbo Zhu[1]
[1]Opus AI Research
[2]Brown University [3]Southeast University [4]University of Toronto [5]Fudan University
`bozheng_li@brown.edu` `wenbo.zhu@berkeley.edu`



Q: What Shot type does the given clip belong to?
A: Over-the-shoulder shot.

Q: Clarify shot motion of given video and explain grounds for the judgment.
A: The shot-motion of the given video is handheld. The camera exhibits slight vertical and horizontal jitters as it follows the girl walking and then moves to focus on the two boys conversing.

Q: what is the cut type used in the given video, and what is its function?
A: The video uses emphasis/deemphasis cut. By cutting from medium shot of reporter, Gayle Hailstorm, to close-up of Heather in car, the emphasis cut draws attention to Heather.

Figure 1. **The overview of our proposed VEU-Bench**. VEU-Bench covers 10 editing dimensions, evaluating models on tasks ranging from recognition to reasoning and judging, providing a robust evaluation of video editing understanding across various aspects and levels of difficulty.

Abstract

*Widely shared videos on the internet are often edited. Recently, although Video Large Language Models (Vid-LLMs) have made great progress in general video understanding tasks, their capabilities in video editing understanding (VEU) tasks remain unexplored. To address this gap, in this paper, we introduce VEU-Bench (**V**ideo **E**diting **U**nderstanding **Bench**mark), a comprehensive benchmark that categorizes video editing components across various dimensions, from intra-frame features like shot size to inter-shot attributes such as cut types and transitions. Unlike previous video editing understanding benchmarks that focus mainly on editing element classification, VEU-Bench encompasses 19 fine-grained tasks across three stages: recognition, reasoning, and judging. To enhance the annotation of VEU automatically, we built an annotation pipeline integrated with an ontology-based knowledge base. Through extensive experiments with 11 state-of-the-art Vid-LLMs, our findings reveal that current Vid-LLMs face significant challenges in VEU tasks, with some performing worse than random choice. To alleviate this issue, we develop Oscars[1], a VEU expert model fine-tuned on the curated VEU-Bench dataset. It outperforms existing open-source Vid-LLMs on VEU-Bench by over 28.3% in accuracy and achieves performance comparable to commercial models like GPT-4o. We also demonstrate that incorporating VEU data significantly enhances the performance of Vid-LLMs on general video understanding benchmarks, with an average improvement of 8.3% across nine reasoning tasks. The code and data are available at project page*

---

[1] Named after the Academy Awards.

1. Introduction

Nowadays large volumes of videos circulating on the internet are edited videos. Video editing involves processing and combining raw footage [29]. Video Editing Understanding (VEU) focuses on identifying and interpreting editing elements in videos, including 'nouns' like shot attributes [1] and 'verbs' such as cuts [4, 32] and transitions [12, 34].

The value of VEU tasks lies in two main areas. First, understanding editing elements enables assisted automatic editing [10, 35], improving efficiency and allowing quality assessment of edited videos [46]. VEU also lowers the barrier for beginners learning video editing. Second, VEU data strengthens the ability to reason abstractly of video models. Editing elements are abstract concepts [8, 29] derived from specialized techniques, not directly present in the real world. For example, understanding a *"Match Cut"* requires recognizing visual patterns, like shape or movement alignment, across scenes. VEU demands knowledge of editing patterns, as such abstractions aren't readily observable in the real world, making VEU highly suitable as abstract reasoning data for video models.

Since VEU benchmarks [1, 32, 34] have primarily focused on classification, the reasoning and interpretation of video editing components remain underexplored. With the recent advancements in Video Large Language Models(Vid-LLMs)[7, 9] capable of performing high-level causal reasoning on video inputs, there is strong potential for utilization of Vid-LLMs in VEU tasks. This makes a more comprehensive and challenging VEU benchmark increasingly indispensable for thoroughly evaluating the VEU ability of Vid-LLMs.

To fill this gap, we propose VEU-Bench (**V**ideo **E**diting **U**nderstanding **Bench**mark), a comprehensive benchmark that categorizes video editing components across various dimensions, from intra-frame features like shot size [1] to inter-shot attributes like cut types [32] and transitions [34]. Unlike previous benchmarks [1, 4, 12, 32, 34] which focus mainly on editing element classification, VEU-Bench covers 19 fine-grained tasks across three stages: recognition, reasoning, and judging, as illustrated in Figure 1.

For benchmark construction, we curate video editing data from previous studies [1, 32, 34] and, guided by professional editing tutorials [29, 42], organize features and functions of editing elements into a detailed knowledge base. Using this knowledge base, we develop an ontology-based annotation pipeline [16] to rewrite abstract editing features, such as *"Vertical camera movement without changing base position"*, into video-specific descriptions like *"The camera tilts vertically upward from the woman's mouth to her eyes, with no lateral movement"*. Similarly, for judging tasks, functions are rewritten in a video-specific context. Through this pipeline, we extend VEU tasks from classification to reasoning and judging, introducing VEU-50K, a well-crafted dataset with around 50k VEU samples. Additionally, the proposed knowledge base provides domain-specific knowledge to aid LLMs in evaluating responses to open-ended questions.

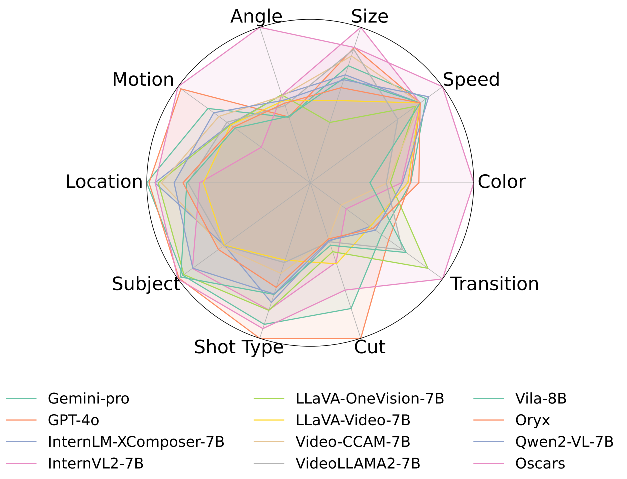


Figure 2. **The performance of 11 Vid-LLMs and the proposed expert model, Oscars, on VEU-Bench.** We normalize the results per dimension for clearer comparisons.

We conduct extensive evaluations of 11 Vid-LLMs on VEU-Bench and found that current state-of-the-art models struggle to understand editing components and patterns, as shown in Figure 2. Therefore, we developed Oscars, an expert Vid-LLM with enhanced video editing comprehension abilities. Trained on VEU-50K training set, Oscars demonstrates superior performance across all VEU tasks, outperforming the leading open-source model LLaVA-Onevision [40] by 28.3% and surpassing Gemini by 4.0%[37], as illustrated in Figure 2 and Table 2. Additionally, Oscars achieves improved performance on several general video understanding benchmarks [11, 20, 24], with average gains of 5.1%, 4.8%, and 6.6% on reasoning-related tasks in Video-MME [11], MVBench [20], and TempCompass [24], demonstrating the value of video editing data in fostering high-level abstract learning, which in turn benefits general video understanding capabilities.

Our contributions are summarized as follows:

- We introduce VEU-Bench, the first comprehensive benchmark designed to evaluate Vid-LLM performance on video editing understanding tasks, covering 19 tasks across 10 dimensions and 3 levels. We also provide 50k high-quality data points to advance research in video editing understanding.
- We conduct a thorough assessment of the current state-of-the-art Vid-LLMs on the proposed VEU-Bench. Our findings reveal that current Vid-LLMs struggle with understanding video editing components, especially in reasoning and judging tasks.

- By fine-tuning on carefully curated VEU-50k dataset, our proposed baseline Vid-LLM, Oscars, surpasses existing state-of-the-art models with a 28.3% overall improvement on VEU-Bench and an average 3% improvement on other general video benchmarks.

2. Related Work

2.1. Video Editing Understanding

Video editing involves complex tasks that go beyond the pixel-level manipulations of video frame [36]. Editing video clips [42] is especially challenging in video understanding, requiring a nuanced grasp of both content and editing principles[8]. Prior studies have explored various technical aspects of editing, including camera shot settings [1, 2, 15, 39], transitions between shots [34], visual effects [12, 46], and cut types [4, 31, 32]. Progress in video editing understanding has been hindered by the lack of large, high-quality public datasets, the subjective nature of editing quality assessment, and the limitations of current video understanding models.

2.2. Video Large Language Model

Recently, the advancement of Vid-LLMs has driven a significant leap forward in the video domain. This progress is highlighted by the introduction of pioneering Vid-LLMs [3, 5–7, 9, 14, 18, 21–23, 27, 40, 44, 48–52]. By integrating a visual encoder and training the projector on large-scale multimodal instruction datasets, these models demonstrate remarkable performance in video understanding tasks. VideoChatGPT [27] and LLaVA-Video [52] utilize LLM and VLM to generate video instruction-tuning data. Works like Qwen2-VL [40], VideoLLaMA2 [7], Video-CCAM [9] and InternLM [51] proposed new design insights for video-llm construction. While these models excel in general video understanding benchmarks, their potential applications in video editing tasks remain underexplored. EditQA-2k [46] explored the capability of Vid-LLMs in analyzing edited video content, but a comprehensive evaluation of Vid-LLMs in understanding video editing components remains absent.

2.3. Video Understanding Benchmark

As Video Large Language Models (Vid-LLMs) advance, various benchmarks [11, 20, 24, 33, 41, 43, 45, 53] have been developed to evaluate their capabilities. Benchmarks like Video-MME [11], MVBench [20], and TempCompass [24] focus on general understanding. Others, such as LongVideoBench [43], MLVU [53], and LVBench [41], examine long-form videos processing ability of Vid-LLM. While these benchmarks offer insights into the general understanding of Vid-LLMs, they overlook complex tasks like video editing understanding that require advanced reasoning. Additionally, most benchmarks are formatted with only multiple-choice QA, making it hard to evaluate the reasoning abilities that require evaluation in an open-ended way. Therefore, we propose VEU-Bench, a comprehensive video editing understanding benchmark to thoroughly evaluate Vid-LLMs' video editing understanding ability with 50K high-quality video editing understanding data.

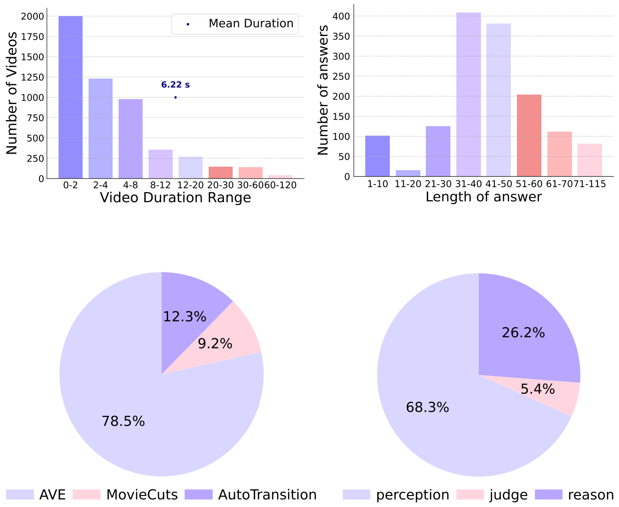


Figure 3. The static of our proposed VEU-Bench.

3. Benchmark

3.1. Overview

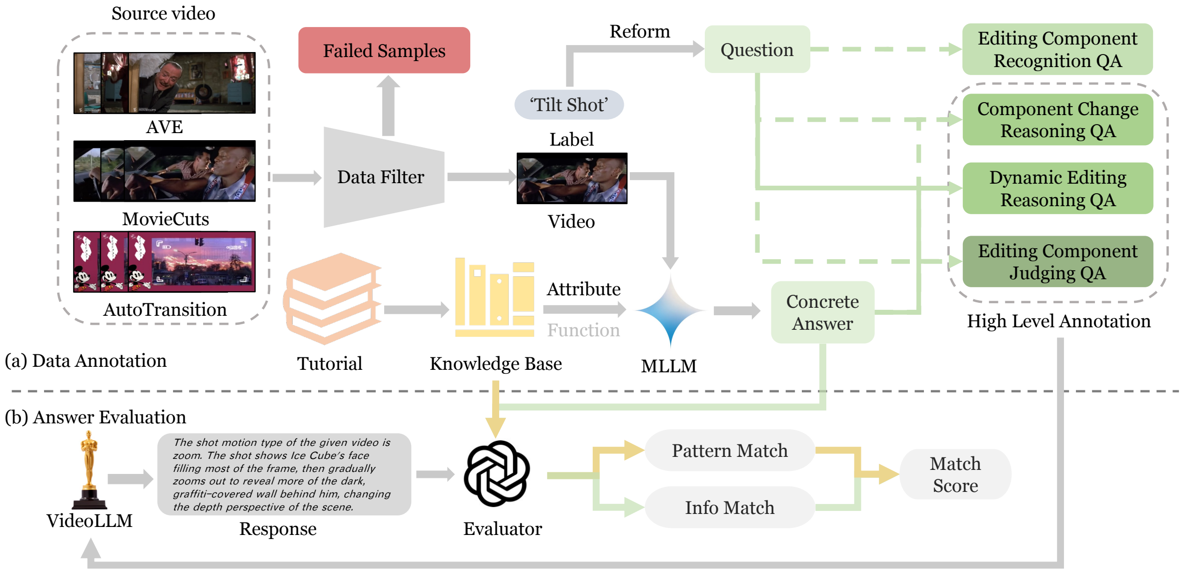
As shown in Figure 3, VEU-Bench encompasses editing components across 10 dimensions and 3 levels, including a total of 30,000 videos and 49,536 QA samples. The training set contains 45,154 samples, and the test set contains 4,382 samples. The video durations range from 1 to over 60 seconds, with the majority between 1 and 12 seconds. We compare our proposed VEU-Bench with previous benchmarks in Table 1. We begin by introducing the task definition of VEU-Bench in Section 3.2, covering both task dimensions and levels. Next, the dataset construction process is detailed in Section 3.3. Finally, we present the evaluation methods in Section 3.4.

3.2. Task Definition

In this section, we provide detailed task definition of VEU-Bench, focusing on two aspects: task dimensions and task levels. Detailed descriptions are attached in the Appendix.

3.2.1. Task Dimensions

Referring to industry definitions [8, 42], we classify video editing components into three categories: *Intra-frame*: Involves only a single frame. *Intra-shot*: Spans multiple frames within a single scene. *Inter-shot*: Covers multiple



The shot motion type of the given video is zoom. The shot shows Ice Cube's face filling most of the frame, then gradually zooms out to reveal more of the dark, graffiti-covered wall behind him, changing the depth perspective of the scene.

Figure 4. **The overview of our data annotation pipeline.** (a) shows the data annotation process for reasoning and judging tasks. Based on an established knowledge base, the annotator selects the most relevant attribute or function and reformulates a video-specific answer to create the QA pair. (b) Indicate evaluation mechanism, the response is matched against the corresponding abstract feature in the knowledge base, as well as compared with the annotated answer to calculate an overall score.

Table 1. **The comparison between VEU-Bench and previous VEU benchmarks.** VEU-Bench encompasses a wider range of video editing components and includes high-level reasoning and judgment tasks.

Dataset	Size	Shot	Cut	Transition	Reasoning	Raw Video
MovieNet [15]	1k	✓	✗	✗	✗	✗
AVE [1]	200k	✓	✗	✗	✗	✗
AutoTransition [34]	30k	✗	✗	✓	✗	✓
MovieCuts [32]	174k	✗	✓	✗	✗	✓
Edit3k [12]	3K	✓	✗	✓	✗	✗
EditedVideo2K [46]	2K	✗	✗	✓	✓	✓
Oscar (ours)	50K	✓	✓	✓	✓	✓

shots across different scenes. Each category includes various editing components, resulting in a total of 10 distinct dimensions as follows:

Intra-frame. (1) *Shot Size:* The proportion of the setting or subject that is visible within the frame of a shot. (2) *Shot Angle:* The angle or perspective from which the camera views the subject, influencing how the subject appears in the frame (e.g., high-angle, low-angle). (3) *Shot Location:* The physical environment or setting in which the scene takes place, providing context or background. (4) *Shot Subject:* The main subject that is highlighted or conveyed within the shot. (5) *Shot Type:* The composition of a shot with respect to the number of featured subjects and their physical relationship to each other as well as to the camera. (6) *Shot color:* The color grading of video clips to create a specific visual tone or mood.

Intra-shot. (7) *Shot motion:* The movement of the camera during the process of taking a shot. (8) *Shot speed:* The playback speed to create different effects within a shot.

Inter-shot. (9) *Cut:* Composed of two adjacent shots and transition between them, without special visual effects and achieved simply by cutting. (10) *Transition:* Moving from one shot to next with special visual effects, creating either an intentional visual impact as one clip changes to another.

3.2.2. Task Levels

While previous VEU benchmarks [1, 32, 34] mainly focus on recognition, we expand the scope of the task to include reasoning and judging, thereby enabling a more comprehensive evaluation. The task levels are constructed as follows:

Recognition. At this level, the model should correctly classify different categories within various task dimensions, in form of multichoice question answering.

Reasoning. At this level, the model is required not only to correctly identify corresponding elements but also to provide evidence and rationale. Two aspects are included in the reasoning task. First, the reasoning of dynamic editing elements e.g. cut and transition. Second, reasoning on changing static editing elements e.g. change of shot size and shot angle. For these tasks, the model must spot and explain changes within a single shot.

Judging. At this level, model is expected to evaluate the functions and benefits of editing elements, like cut and shot types, by interpreting their role within video based on both the content and the creator's intent. Judging assesses model's ability to applying its understanding to explain how each editing choice contributes to storytelling or desired impact within a real editing context.

3.3. Dataset Construction

3.3.1. Video Collection

We begin by collecting video data from existing video editing datasets: AVE [1], MovieCuts [32], and AutoTransition [34]. We focus on short clips to enable Vid-LLMs to process fundamental units effectively and ensure compatibility within context length limitations [6]. We exclude unsuitable dimensions, filter out incorrect annotations using Gemini-1.5-pro [37], and balance data distribution to mitigate the long-tail effect from dominant labels.

3.3.2. Automatic Annotation

Recognition. In this task, we introduce two new dimensions: shot color and shot speed. Shot color labels are generated based on HSV color thresholds [17], while shot at different speeds is created using ffmpeg [38] to apply speed adjustments, including both fast and slow motion effects. Due to broad variety of transition types, transition recognition task randomly samples three other categories per sample to structure multiple-choice questions. For other recognition tasks, we include all dimensions in question options.

Reasoning & Judging. Reasoning and judging video editing components require costly manual annotation due to the need for expert knowledge. Direct annotation by MLLM also proves challenging, as evidenced in Figure 2, even top-performing models like Gemini[37] struggle with VEU tasks. To overcome this hindrance, we propose an automatic pipeline that leverages the ontology-based system[16] grounded in an abstract knowledge base. As shown in Figure 4, for each editing component, we construct a professional knowledge base by integrating definitions from existing video editing tutorial [42]. Specially, for reasoning tasks, we define "key attributes" describing abstract patterns of each dimension. Each dimension contains features that serve as accordance for reasoning and identification. For the Judging task, we define the functions of editing elements within video content, ensuring video-independent applicability. During annotation, MLLM selects the most relevant attributes or functions based on the video content, and then abstract terms like "object" or "scene" in pre-defined features are replaced with specific terms to ensure content relevance. For example, attributes for *"match cut"* include elements like *"connecting two similarly shaped objects across frames."* While annotating a particular video, this feature will be rewritten as *"a match cut connects a similarly shaped bone and spaceship across frames."* This annotation process simplifies open-ended reasoning to a rewriting task, enables our benchmark to deliver high-quality video editing annotations with minimal human involvement and guarantees LLM annotators to generate well-crafted responses with less need for deep video editing knowledge. After completing the construction of the dataset, we conducted manual reviews and user studies to further validate the data quality. Details can be found in the appendix.

3.4. Evaluation

3.4.1. Metrics

For the Reasoning and Judging tasks, which require open-ended answers, we adopt an approach inspired by Video-ChatGPT [28]. We leverage GPT-4 [30] to assign relative scores to generated predictions. Directly scoring responses against correct answers, however, may lead to inflated scores, as responses often contain correct descriptive information but misinterpret editing patterns. To address this, we propose a scoring system that incorporates pattern-matching regularization based on a knowledge base constructed for dataset annotation, as shown in Figure 4 (b).

Specifically, given an editing component ontology O (attributes for reasoning tasks and functions for judging tasks), ground-truth answerA_{gt}, and predicted answer A_{pd}, matching score between response and answer is defined as:

$$S_{\text{match}} = \frac{PM(A_{\text{pd}}, O) + IM(A_{\text{pd}}, A_{\text{gt}})}{2} \in [0, 5]$$

where PM (pattern matching score) measures how well A_{pd} aligns with the editing pattern ontology O, focusing on the video-agnostic editing pattern. IM (information matching score) assesses the alignment of specific objects and visual details between A_{pd} and A_{gt}. Additionally, the accuracy Acc of the editing component is computed by the LLM for open-ended questions. To evaluate performance on reasoning and judging tasks, we scale accuracy to a 0-5 range and combine it with the matching score, resulting in:

$$S_{oe} = \frac{(5 \times Acc + S_{\text{match}})}{2} \in [0, 5]$$

For Recognition tasks, we use accuracy as the metric, scaling it to the 0-5 range in sync with the reasoning and judging metrics, defined as $S_{mc} = 5 \times Acc \in [0, 5]$.

3.4.2. Prompt Set

Video editing understanding requires domain-specific knowledge, therefore, a simple prompt that directly queries the model cannot thoroughly evaluate the VEU ability of Vid-LLMs. Consequently, we design two extra prompts to optimize model performance to achieve the best results. First, we incorporate **context prompt** and integrate detailed definitions of editing components into the prompt to help the model better grasp the concept of these components and focus on understanding the video content. Second, **guidance prompt** that underscores general video task requirements are included. Also, we adapt different guidance

Table 2. **The performance of current state-of-the-art Vid-LLMs and the proposed Oscars on VEU-Bench.** The best performance is highlighted in bold and the second-best is underlined. Notably, Oscars surpasses the base model Qwen2-VL-7B by 39.6% and outperforms Gemini-1.5-pro by 4%, achieving performance comparable to GPT-4o.

Level	**Dimension**	Gemini-1.5-Pro [37]	GPT-4o [30]	InternLM-X-7B [51]	InternVL2-8B [6]	LLaVA-OV-7B [19]	LLaVA-Video-7B [52]	Video-CCAM-7B [9]	VideoLLaMA2-7B [7]	VILA-8B [47]	Oryx-7B [25]	Qwen2-VL-7B [40]	**Oscars-7B(ours)**
Frame Numbers		1FPS	8	64	16	16	16	96	32	16	32	1FPS	1FPS
Recognition	Shot Subject	77.3	79.8	52.2	71.5	76.8	52.2	53.3	52.5	78.7	55.8	71.3	**80.4** (+9.1)
	Shot Color	49.3	49.1	48.0	44.7	39.3	48.0	38.0	37.3	29.3	**53.3**	45.3	51.3 (+6.0)
	Shot Size	55.6	63.3	53.0	51.5	22.4	43.3	52.0	50.7	49.3	51.5	51.7	**66.2** (+14.5)
	Shot Angle	54.5	**60.2**	36.0	36.7	39.8	36.0	39.2	33.4	39.2	36.8	45.8	48.6 (+2.8)
	Shot Location	**90.2**	89.1	88.4	89.5	84.9	88.4	87.7	84.2	87.7	76.5	86.7	90.0 (+3.3)
	Shot Type	76.4	**83.0**	66.4	72.7	71.1	66.4	59.2	55.2	62.7	60.6	62.2	73.7 (+11.5)
	Shot Motion	**36.3**	35.2	30.2	22.9	20.6	30.2	24.0	21.4	18.1	19.5	27.0	34.1 (+7.1)
	Shot Speed	35.1	33.3	33.3	33.3	32.0	33.3	33.3	26.6	33.3	33.3	36.0	**40.1** (+6.8)
	Transition	35.3	51.1	25.4	15.1	49.8	25.3	13.0	38.9	40.5	26.6	27.5	**55.9** (+29.7)
	Cut Type	32.1	**42.4**	17.1	21.4	17.6	17.1	16.7	17.1	17.6	17.6	16.2	28.7 (+12.5)
Score$_{mc}$		2.71	**2.93**	2.25	2.29	2.27	2.20	2.08	2.09	2.28	2.20	2.33	2.85 (+0.52)
Reasoning	Shot Size	1.05	1.28	0.88	2.52	0.86	0.54	1.55	**1.84**	0.97	0.55	0.80	1.13 (+0.16)
	Shot Angle	1.30	1.03	0.90	0.85	0.71	0.72	0.70	0.70	0.08	0.60	0.24	**2.34** (+2.10)
	Shot Location	**3.96**	3.91	3.63	3.27	3.64	1.11	3.34	2.12	2.02	2.78	2.73	3.53 (+1.50)
	Shot Type	3.46	**3.85**	2.84	2.90	3.00	0.66	1.70	1.33	2.60	2.35	2.61	3.81 (+1.07)
	Shot Motion	0.99	1.78	0.68	0.19	1.15	0.67	1.29	1.20	1.15	1.12	1.29	**1.92** (+0.77)
	Transition	**1.26**	0.77	0.19	0.24	0.72	0.11	0.35	0.35	0.30	0.14	0.21	0.91 (+0.70)
	Cut Type	1.94	**2.25**	0.78	1.16	1.05	1.43	0.75	0.81	0.87	0.70	0.80	1.58 (+0.78)
Judging	Shot Type	3.65	**3.95**	3.04	2.97	3.15	0.96	1.42	1.65	2.36	2.08	2.66	3.74 (+1.08)
	Cut Type	1.98	**2.39**	0.34	0.19	0.94	0.97	0.32	0.91	0.51	0.68	0.41	2.04 (+1.63)
Score$_{oe}$		2.11	2.36	1.48	1.59	1.69	0.80	1.27	1.21	1.21	1.22	1.31	**2.23** (+1.17)
Score$_{all}$		2.44	**2.64**	1.79	1.94	1.98	1.50	1.68	1.65	1.75	1.71	1.82	2.54 (+0.72)

prompts for each model while maintaining prompt semantics. Prompts samples are included in the Appendix.

4. Experiments

4.1. Implementation Details

We conduct extensive benchmarking across mainstream Vid-LLMs, including open-source models LLaVA-Video [52], InternLM-X [51], InternVL [6], LLaVA-OneVision [19], Qwen2-VL [40], Video-CCAM [9], VideoLLaMA2 [7], Oryx [25], ViLA [22] and proprietary models include Gemini-1.5-Pro [37] and GPT-4o [30].

For model training, we use Qwen2-VL-7B [40] as the base model and apply LoRA [13] fine-tuning with $r = 16$ and $\alpha = 32$. The learning rate is set to $1e-4$, weight decay to 0.01, and warmup ratio to 0.05. The model is optimized using AdamW [26]. For each video, frames are sampled at 1 fps with a maximum limit of 64 frames. All experiments are conducted on 4 A100 GPUs.

4.2. Main Results

4.2.1. Quantitative Results

We evaluate VEU-Bench on 11 state-of-the-art Vid-LLMs, as shown in Table 2. Current Vid-LLMs exhibit poor performance across all benchmark dimensions, indicating a lack of capability in recognizing and reasoning about video editing components.

In the recognition task, Vid-LLMs perform significantly better at identifying intra-frame editing components than inter-frame and inter-shot content. This is likely due to the simplicity of analyzing a single frame when recognizing intra-frame elements, which does not involve dynamic content changes. Furthermore, in the areas of shot speed and shot motion, models like VideoLLaMA2 perform even worse than random guessing. This is attributed to the use of a uniform sampling strategy, which limits their ability to effectively perceive the concept of speed.

Current Vid-LLMs also show limited performance in reasoning and judging tasks, with an average score below 2 out of 5. Compared to the recognition level, involving changes makes intra-frame features harder to recognize. Additionally, they score lower on cut and transition aspects compared to recognition tasks. Most open-source models exhibit lower performance on judging tasks compared to reasoning tasks in the cut-type dimension. This may be attributed to the challenge of establishing connections between the function of editing elements and specific video content, which goes beyond simple pattern recognition.

Table 3. **The performance of Oscars on representative dimension of general video understanding benchmarks.** Green numbers indicate improvement of Oscars compared to Qwen2-VL-7B.

Benchmark	Dimension	Qwen2-VL-7B	Oscars
VideoMME$_{short}$ [11]	Attribute Perception	73.0	80.3(+7.3)
	Action Reasoning	72.3	76.6(+4.3)
	Information Synopsis	82.9	86.6(+3.7)
MVBench [20]	Unexpected Action	72.0	78.0 (+6.0)
	State Change	47.0	52.5 (+5.5)
	Moving Direction	43.0	46.0 (+3.0)
TempCompass [24]	Order	54.1	62.6 (+8.5)
	Direction	39.7	46.7 (+7.0)
	Speed	43.4	47.6 (+4.2)

Table 4. **The concept experiment on Vid-LLMs.**

Model	**Intra-Frame**	**Intra-Shot**	**Inter-Shot**
Gemini-1.5-Pro [37]	2.76(±0.006)	2.82(±0.006)	2.82(±0.002)
Qwen2-VL [40]	2.63(±0.010)	2.62(±0.009)	2.74(±0.005)
VideoLLaMA2 [7]	2.52(±0.021)	2.54(±0.014)	2.57(±0.009)
LLaVA-Video [47]	2.65(±0.015)	2.56(±0.012)	2.62(±0.015)

In contrast, by training on VEU-50k, our expert model Oscars exhibits improvements across all dimensions compared to Qwen2-VL[40], with gains of 22.3% in Score$_{mc}$ and 89.3% in Score$_{oe}$. Meanwhile, Oscars gains 28.3% higher performance in Score$_{all}$ compared to SOTA open-source model LLaVA-OneVision[19]. It also achieves performance comparable to state-of-the-art commercial models GPT-4o and surpasses Gemini-1.5-pro with 4%. In the more challenging dimensions cut and transition, Oscars demonstrates significantly better performance than the open-source Vid-LLM state-of-the-art, with improvements of 12.2% and 10.9% on Score$_{mc}$ and Score$_{oe}$ respectively, showing exceeding video editing understanding ability. For a more detailed discussion on the dimension-wise performance of Vid-LLMs, please refer to the Appendix.

4.2.2. Qualitative Results

We present a qualitative comparison among our model, Oscars, Qwen2-VL[40] and GPT-4o[30] in Figure 6. Oscars demonstrate superior video editing understanding, accurately identifying the *"Wide"* shot size and function of the *"Smash cut"* cut type. In contrast, Qwen2-vl and GPT-4o confused the nuanced difference between Medium and Wide shot size, and GPT-4o merely recognizes the content visible in the video and lacks the ability to reasonably infer the functionality of editing components. Additional visualization examples can be found in the Appendix.

4.3. Deep Analysis

4.3.1. Results on General Video Benchmarks

We evaluate the performance of Oscars on general video understanding benchmarks, as shown in Table 3. The results indicate that only by fine-tuning on 50k VEU-Bench data, the general video content understanding ability of the model is significantly enhanced. Specifically, Oscars achieves 7.3% improvement on the attribute perception task in VideoMME-short[11], 5.5% improvement on the state change task in MVBench[20] and 8.5% improvement on the Ordering task in TempCompass[24]. Given that VEU-50k contains domain-specific data with low similarity to general video understanding tasks, we attribute this improvement to the rich dynamic content and comprehension demands in the VEU data, which strengthen the temporal understanding ability of Vid-LLMs, leading to better performance on general tasks. We include full result on general video benchmarks in appendix

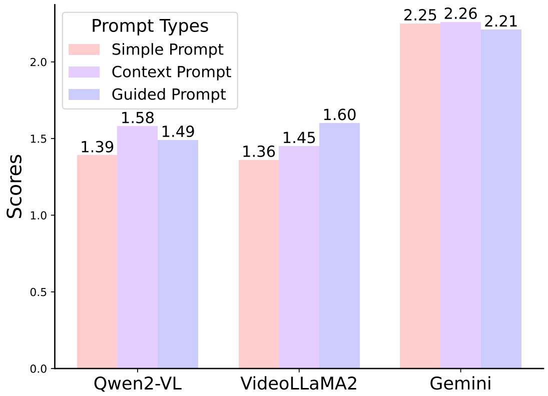


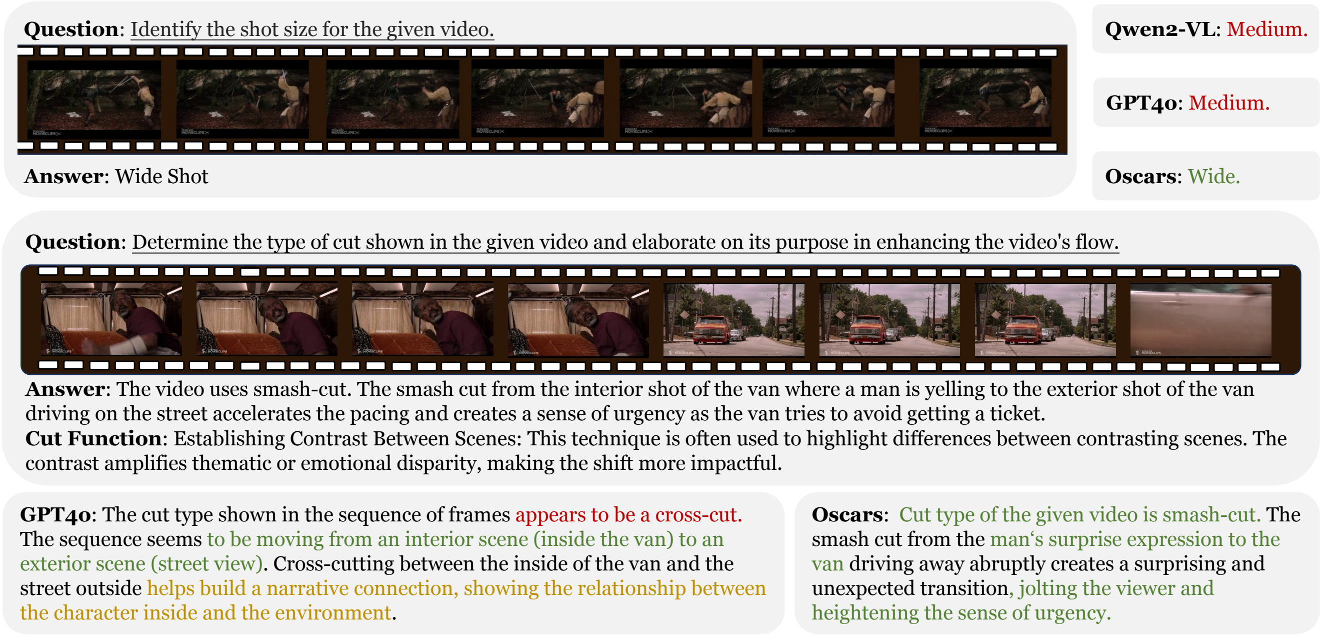
Figure 5. **Ablation results of prompt designs.**We conduct experiments on Qwen2-VL-7B, VideoLLaMA2-7B and Gemini-1.5-Pro.

4.3.2. Different Prompt Designs

We explore the effects of three types prompt designs as discussed in Section 3.4: simple prompt, context prompt, and guided prompt. As shown in Figure 5, compared to simple prompt, the context prompt helps activate editing-related knowledge in LLMs, enhancing the model's ability to understand video editing elements. This leads to 6.6% improvement on VideoLLaMA2 [7] and 13.7% improvement on Qwen2-VL [40]. Due to strong video understanding capabilities of Gemini-1.5-pro[37], the effect of context prompt is less pronounced. Guidance prompts are more beneficial for models with weaker general video understanding capabilities, providing a 10% improvement for VideoLLaMA. For Qwen2-VL-7B and Gemini-1.5-pro, we attribute the performance drop to increased context length and overly restrictive prompts, which limit the models' capabilities.

4.3.3. Exploration of Intrinsic Knowledge

Given two challenges in VEU-Bench tasks: (1) mastering knowledge related to video editing, and (2) accurately identifying video edits, we aim to explore whether the poor performance of video language models (Vid-LLMs) is due to



Question: Identify the shot size for the given video.

Answer: Wide Shot

Qwen2-VL: Medium.

GPT4o: Medium.

Oscars: Wide.

Question: Determine the type of cut shown in the given video and elaborate on its purpose in enhancing the video's flow.

Answer: The video uses smash-cut. The smash cut from the interior shot of the van where a man is yelling to the exterior shot of the van driving on the street accelerates the pacing and creates a sense of urgency as the van tries to avoid getting a ticket.
Cut Function: Establishing Contrast Between Scenes: This technique is often used to highlight differences between contrasting scenes. The contrast amplifies thematic or emotional disparity, making the shift more impactful.

GPT4o: The cut type shown in the sequence of frames appears to be a cross-cut. The sequence seems to be moving from an interior scene (inside the van) to an exterior scene (street view). Cross-cutting between the inside of the van and the street outside helps build a narrative connection, showing the relationship between the character inside and the environment.

Oscars: Cut type of the given video is smash-cut. The smash cut from the man‘s surprise expression to the van driving away abruptly creates a surprising and unexpected transition, jolting the viewer and heightening the sense of urgency.

Figure 6. **The qualitative results of Oscars, GPT-4o and Qwen2-VL-7B on VEU-bench.** Green text indicates the correctly answered part; Red text indicates wrong information; Brown text indicates non-relevant judging on the effect of cut type.

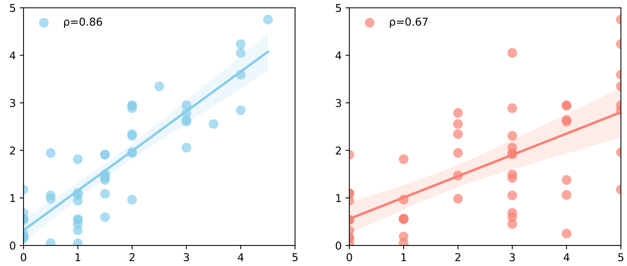


Figure 7. **The alignment between LLM (vertical axis) and human scores (horizontal axis).** Left figure indicates scoring with both IM and PM and right figure indicates scoring with only IM.

a lack of built-in understanding of editing concepts. We design 37 questions covering elements within these dimensions to assess the model's comprehension of editing concepts. The generated answers were evaluated by 10 volunteers, who scored each response from 1 (poor) to 3 (good). As shown in Table 4, both the best- and worst-performing models demonstrated accurate descriptions of editing concepts. The results suggest that the poor performance of Vid-LLMs may be due to weak alignment between the intrinsic knowledge of the language model and the visual perception component. More details can be found in the Appendix.

4.3.4. Scoring System Evaluation

We evaluate the alignment between our proposed scoring system and human preferences to assess the impact of pattern matching regularization. Ten human volunteers rated 50 answer-response pairs sampled from reasoning and judging tasks. As shown in Figure 7, fewer data points concentrated below the diagonal and a higher Spearman coefficient indicate that incorporating pattern matching reduces the bias of the LLM evaluator toward scoring higher based solely on visual factual content. This results in scores that align better with human evaluations, demonstrating the robustness of our proposed evaluation system.

5. Conclusion

In this work, we introduce VEU-Bench, a comprehensive benchmark for evaluating Vid-LLM in the Video Editing Understanding task. Our benchmark analysis reveals significant limitations in the ability of current SOTA Vid-LLMs to comprehend and reason about video editing components. To address this, we present Oscars, a fine-tuned Vid-LLM that demonstrates substantial performance gains on VEU-Bench, exceeding leading open-sourced Vid-LLM by 28.3% and achieving comparable performance with commercial models like GPT-4o and Gemini-pro. Furthermore, Oscars exhibits improved performance on general video understanding benchmarks, which underscores the value of VEU data to serve as the data source for enhancing the abstract reasoning ability of Vid-LLMs. VEU-Bench and Oscars provide valuable resources for advancing research in video editing understanding and enhancing the reasoning capabilities of Vid-LLMs.

References

[1] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In European Conference on Computer Vision, pages 201–218. Springer, 2022. 2, 3, 4, 5

[2] Noël Burch. Theory of film practice. Princeton University Press, 2014. 3

[3] Jiawang Cao, Yongliang Wu, Weiheng Chi, Wenbo Zhu, Ziyue Su, and Jay Wu. Reframe anything: Llm agent for open world video reframing. arXiv preprint arXiv:2403.06070, 2024. 3

[4] Boris Chen, Amir Ziai, Rebecca S Tucker, and Yuchen Xie. Match cutting: Finding cuts with smooth visual transitions. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2115–2125, 2023. 2, 3

[5] 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. 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, 6

[7] 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. 2, 3, 6, 7

[8] Ken Dancyger. The technique of film and video editing: history, theory, and practice. Routledge, 2018. 2, 3

[9] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. arXiv preprint arXiv:2408.14023, 2024. 2, 3, 6

[10] Nathan Frey, Peggy Chi, Weilong Yang, and Irfan Essa. Automatic non-linear video editing transfer. arXiv preprint arXiv:2105.06988, 2021. 2

[11] 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. 2, 3, 7

[12] Xin Gu, Libo Zhang, Fan Chen, Longyin Wen, Yufei Wang, Tiejian Luo, and Sijie Zhu. Edit3k: Universal representation learning for video editing components. arXiv preprint arXiv:2403.16048, 2024. 2, 3, 4

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. 6

[14] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14271–14280, 2024. 3

[15] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pages 709–727. Springer, 2020. 3, 4

[16] Khushboo Khurana and MB Chandak. Study of various video annotation techniques. International Journal of Advanced Research in Computer and Communication Engineering, 2(1):909–914, 2013. 2, 5

[17] Suzi Kim and Sunghee Choi. Automatic color scheme extraction from movies. In Proceedings of the 2020 international conference on multimedia retrieval, pages 154–163, 2020. 5

[18] Bozheng Li, Mushui Liu, Gaoang Wang, and Yunlong Yu. Frame order matters: A temporal sequence-aware model for few-shot action recognition. arXiv preprint arXiv:2408.12475, 2024. 3

[19] 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. 6, 7

[20] 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. 2, 3, 7

[21] 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. 3

[22] 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. 6

[23] Mushui Liu, Bozheng Li, and Yunlong Yu. Omniclip: Adapting clip for video recognition with spatial-temporal omni-scale feature learning. arXiv preprint arXiv:2408.06158, 2024. 3

[24] 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, 7

[25] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. arXiv preprint arXiv:2409.12961, 2024. 6

[26] I Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 6

[27] 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. 3

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 2024. 5

[29] Christian Metz. Film language: A semiotics of the cinema. University of Chicago Press, 1991. 2

[30] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 5, 6, 7

[31] Alejandro Pardo, Fabian Caba, Juan León Alcázar, Ali K Thabet, and Bernard Ghanem. Learning to cut by watching movies. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6858–6868, 2021. 3

[32] Alejandro Pardo, Fabian Caba Heilbron, Juan León Alcázar, Ali Thabet, and Bernard Ghanem. Moviecuts: A new dataset and benchmark for cut type recognition. In European Conference on Computer Vision, pages 668–685. Springer, 2022. 2, 3, 4, 5

[33] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems, 36, 2024. 3

[34] Yaojie Shen, Libo Zhang, Kai Xu, and Xiaojie Jin. Autotransition: Learning to recommend video transition effects. In European Conference on Computer Vision, pages 285–300. Springer, 2022. 2, 3, 4, 5

[35] John R Smith, Dhiraj Joshi, Benoit Huet, Winston Hsu, and Jozef Cota. Harnessing ai for augmenting creativity: Application to movie trailer creation. In Proceedings of the 25th ACM international conference on Multimedia, pages 1799–1808, 2017. 2

[36] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. Diffusion model-based video editing: A survey. arXiv preprint arXiv:2407.07111, 2024. 3

[37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. 2, 5, 6, 7

[38] Suramya Tomar. Converting video formats with ffmpeg. Linux journal, 2006(146):10, 2006. 5

[39] Bartolomeo Vacchetti and Tania Cerquitelli. Movie lens: Discovering and characterizing editing patterns in the analysis of short movie sequences. In European Conference on Computer Vision, pages 660–675. Springer, 2022. 3

[40] 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. 2, 3, 6, 7

[41] 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. 3

[42] Michael Wohl. Editing techniques with final cut pro. Peachpit Press, 2002. 2, 3, 5

[43] 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. 3

[44] Yongliang Wu, Bozheng Li, Jiawang Cao, Wenbo Zhu, Yi Lu, Weiheng Chi, Chuyun Xie, Haolin Zheng, Ziyue Su, Jay Wu, et al. Zero-shot long-form video understanding through screenplay. arXiv preprint arXiv:2406.17309, 2024. 3

[45] Yongliang Wu, Wenbo Zhu, Jiawang Cao, Yi Lu, Bozheng Li, Weiheng Chi, Zihan Qiu, Lirian Su, Haolin Zheng, Jay Wu, et al. Video repurposing from user generated content: A large-scale dataset and benchmark. arXiv preprint arXiv:2412.08879, 2024. 3

[46] Lu Xu, Sijie Zhu, Chunyuan Li, Chia-Wen Kuo, Fan Chen, Xinyao Wang, Guang Chen, Dawei Du, Ye Yuan, and Longyin Wen. Beyond raw videos: Understanding edited videos with large multimodal model. arXiv preprint arXiv:2406.10484, 2024. 2, 3, 4

[47] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. arXiv preprint arXiv:2408.10188, 2024. 6, 7

[48] 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. 3

[49] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840, 2024.

[50] 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.

[51] 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. 3, 6

[52] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713, 2024. 3, 6

[53] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv preprint arXiv:2406.04264, 2024. 3