

# VEU-Bench: Towards Comprehensive Understanding of Video Editing

## Supplementary Material

In this supplementary material, we provide a detailed description of the dataset construction process, an in-depth analysis of model performance on VEU-Bench as well as the impact of training data on model performance. Section A details the dataset filtering and splitting processes, along with the specific distribution within VEU-Bench. Section B presents examples of QA pairs and the details of knowledge base construction. Section C provides prompts used for annotation, inference, and evaluation. Section D discusses the impact of different proportions of task types on model training. Section E elaborates on the details of the concept experiments. Section F.1 analyzes the model performance on VEU-Bench with a category-wise breakdown. Section G showcases visual results and comparisons for each task, while Section H discusses the limitations of VEU-Bench and outlines directions for future research.

## A. Dataset Construction Details

### A.1. Curation and Filter

**Video-wise:** We curated videos from AVE[1], MovieCuts[29], and AutoTransition[31], trimming them into single-shot clips where each clip contains only one editing component. This ensures stable model outputs and focuses recognition on a single type of component. We filtered out videos shorter than 0.5 seconds to ensure sufficient temporal information for recognizing stable editing features. For transitions that occur within a single frame, we trimmed 1-second clips, including 0.5 seconds before and 0.5 seconds after the transition.

**Dimension-wise:** We removed incorrect dimensions, such as "inter-shot" in shot type, which cannot be determined through a single clip. For instance, AVE[1] only annotates the insert shot itself without including context clips, contradicting the definition of inter-shot. Due to the video understanding limitations of current Vid-LLMs, most of which cannot handle multi-video inputs, we focused on single continuous video clips, which can include either multi-shot or single-shot sequences. We also removed videos with "other" labels to ensure a well-defined and clean benchmark.

**Label-wise:** In the AVE[1] dataset, there are multi-labeled videos. We transformed some of these into change-detection tasks, verifying labels through Gemini[34] and GPT[27]. Specifically, for change-detection tasks, we sampled the first two frames and the last two frames, requiring both Gemini and GPT to recognize the shot attributes of the changing tasks. Only videos where the labels matched the answers from both GPT and Gemini were retained.

For shot-size changes, we only kept evident changes, such as from extreme-wide to medium, and removed subtle changes, such as from extreme-wide to wide. For other tasks, we balanced the label distribution across dimensions in the test set. As shown in Table, the original annotations of the AVE and MovieCuts datasets exhibited significant imbalance. We balanced the number of labels to provide more reasonable and effective evaluations, avoiding potential model bias towards single-category answers that could artificially boost accuracy.

### A.2. Split of evaluation data

For MovieCuts[29] and AutoTransition[31], we follow the train and test split from the original datasets and subsample videos from the test set to construction evaluation data of VEU-Bench. For AVE[1], since no split of the datasets is publicly available, we split the AVE datasets with the same portion according to the original paper[1] and will make a split of datasets publicly available.

### A.3. Detailed category distribution

We present the category-wise data distribution of the VEU-Bench evaluation set in Figure A.1 and compare it with the original distributions from AVE [1], AutoTransition [31], and MovieCuts [29]. The imbalance in the original datasets has been significantly reduced, enabling a more stable test setup and trustworthy results, while preventing Vid-LLMs from achieving higher scores by focusing on a single dominant category.

## B. Annotation

**Question List:** We provide the question list of each task in Table 5. We introduce four questions per task to guarantee diverse input form and stable and general output of the model. **Knowledge Base Construction:** Details about key attributions and functions can refer to the knowledge base JSON files we attached in the supplementary material. For the knowledge construction process, except for transition which has more than 70 types of transition and has no detailed and formal definition of each type, we collect definitions and functions of each component from the lecture notebook and polish the definition into detailed, non-overlapped definitions and functions. For the transition effect knowledge base, we sampled one video from each transition type and asked Gemini to give an abstract definition of the transition type based on the prompt in Figure C.1.

**Automatic annotation** Each video in the reasoning and judging tasks are rewritten into a video-specific answer

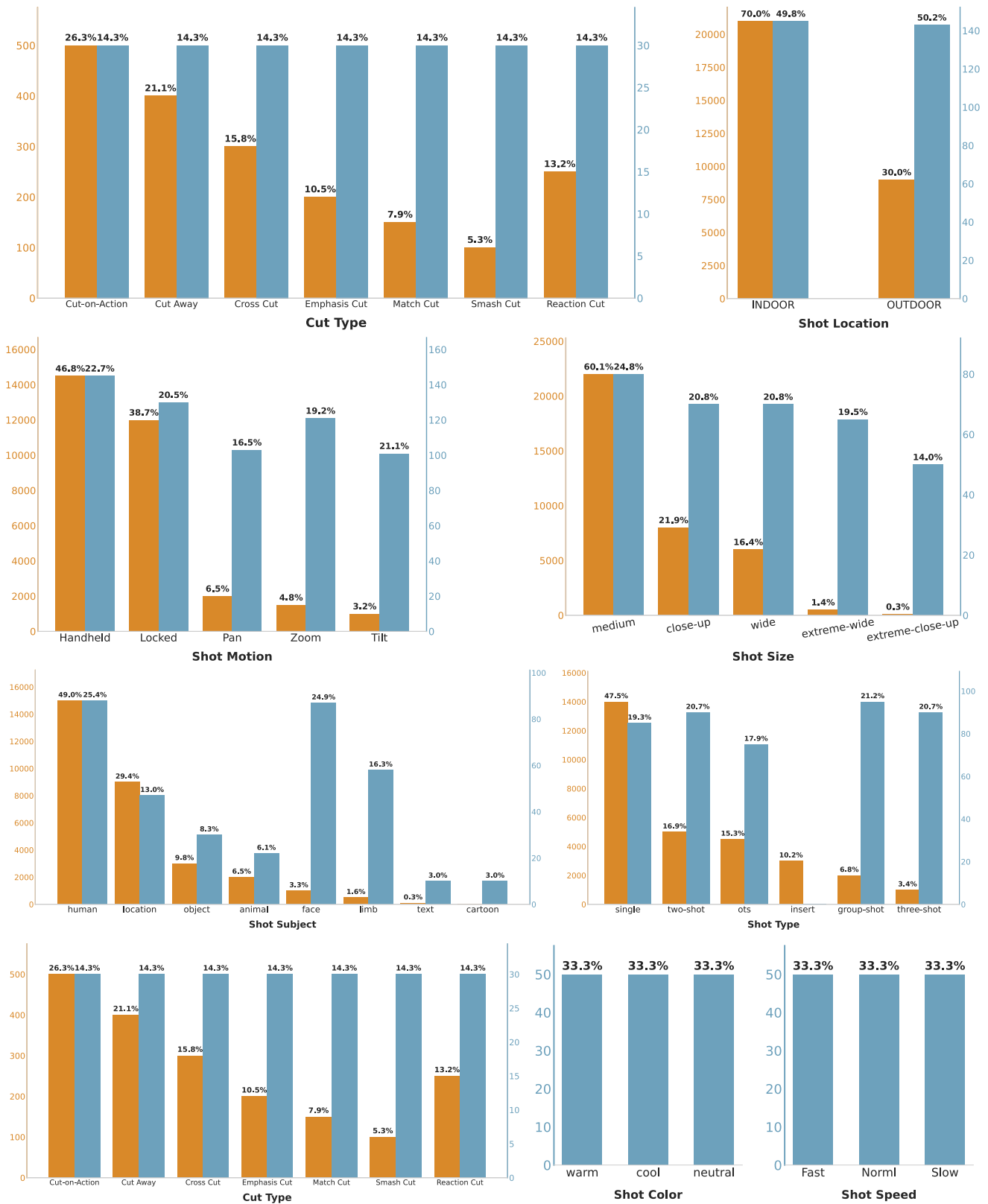


Figure A.1. Comparison between the original data distribution (sky blue) and the VEU-Bench evaluation set distribution (deep blue). Zoom in to get the best reading experience.

Table 5. Questions List of VEU-Bench Across 19 tasks.

Dimension	Questions
<i>Editing Dimension = Shot Color, Shot Speed, Shot Size, Shot Angle, Shot Type, Shot Motion, Shot Location, Shot Subject, Cut Type, Transition</i>	
Recognition	(1) To which { <i>Editing Dimension</i> } does the given video belong; (2) Determine the { <i>Editing Dimension</i> } used in the given video; (3) What category of { <i>Editing Dimension</i> } does the given video represent; (4) Identify the { <i>Editing Dimension</i> } for the given video.
<i>Editing Dimension = Shot Size, Shot Angle, Shot Location</i>	
Change Reasoning	(1) In the given video, identify the { <i>Editing Dimension</i> } change from one type to another and state the basis for your determination; (2) Based on the given video, specify the { <i>Editing Dimension</i> } transition from one type to another and provide the criteria for this judgment; (3) In the given video, clarify the { <i>Editing Dimension</i> } shift from one type to another and explain the clue for this change; (4) For the given video, outline the change in { <i>Editing Dimension</i> } from one type to another and explain the basis for your decision.
<i>Editing Dimension = Shot Motion, Shot Type, Cut Type, Transition</i>	
Dynamic Reasoning	(1) In the given video, clarify the { <i>Editing Dimension</i> } and explain the grounds for your judgment. (2) For the given video, outline the { <i>Editing Dimension</i> } type and state the basis for your determination; (3) Based on the given video, specify the { <i>Editing Dimension</i> } and provide the criteria for this decision; (4) Identify the { <i>Editing Dimension</i> } of the given video and explain the basis for your choice.
<i>Editing Dimension = Shot Type, Cut Type</i>	
Judging	(1) Based on the given video, what is the { <i>Editing Dimension</i> } used in the given video, and what is its function? (2) Based on the given video, what kind of { <i>Editing Dimension</i> } do you observe in the given video, and what function does it serve; (3) Determine the type of { <i>Editing Dimension</i> } shown in the given video and elaborate on its purpose in enhancing the video’s flow; (4) Based on the given video, Decide the { <i>Editing Dimension</i> } used in the vide, and what role does it play in the given video?

based on the prompt provided in Section C.

### C. Prompt

In this section, we present the prompts covering three aspects: annotation prompt in Figure C.1, task inference prompt in Figure C.2, and result evaluation prompt in Figure C.3.

### D. VEU Data as instruction tuning data

We investigate the effect of data from different-level tasks serving as instruction tuning data on VEU-Bench. As shown in Table 6. We have the following observations. (1) The coverage of dimensions in the training data is crucial. While judging data involves more complex reasoning tasks, its dimensional limitations result in a 7% gap in overall performance compared to recognition data, which covers all dimensions. (2) Reasoning data encompasses more comprehensive dimensions and serves as an upgrade of recognition tasks. Therefore, under the same amount of single-task training data, reasoning data achieves the best performance on VEU-Bench, 4.5% higher than recognition data and 11.0% higher than judging data. (3) By leveraging simple recognition tasks, the model gains great improve-

Table 6. The impact of different task type proportions in training data on model performance. Due to complexity and diversity, data of **reasoning tasks** lead to best performance.

Rec(%)	Rea(%)	Jud(%)	Rec	Reason	Jud	Overall
0	0	0	2.33	1.24	1.54	<b>1.82</b>
100	0	0	3.15	1.51	2.51	<b>2.44</b>
0	100	0	2.70	2.24	2.75	<b>2.53</b>
0	0	100	2.55	1.74	2.99	<b>2.28</b>
50	50	0	3.05	1.92	1.91	<b>2.48</b>
0	50	50	2.60	1.90	2.50	<b>2.32</b>
50	0	50	3.00	1.51	2.34	<b>2.35</b>
33	33	33	2.95	1.83	2.47	<b>2.46</b>

ment compared to base model Qwen2-VL[37]. (4) Regarding task dimensions, the data for each task type has the strongest impact on improving performance within its respective task while reasoning data shows the best effect in enhancing performance across other tasks.

In future work, we will further expand judging-type data to achieve training datasets that balance both diversity and complexity.



Figure C.1. Annotation prompt we used in the dataset generation stage.

## E. Details of Concept Experiment

In Table 7, we provide the 10 question samples from the concept experiment, covering all dimensions within video editing understanding and listing answers from the best model and worst model on the benchmark. It is evident that all these models have shown good responses and got high scores from human judgment as shown in Table 4. Volunteers evaluate the answer based on the following criteria: “Given the definition of an editing element and the model’s response regarding that element, evaluate whether the model has accurate knowledge of the specified editing

element and assign a score based on the degree of alignment. **Good (3):** The model provides a comprehensive definition that aligns closely with the given answer. **Middle (2):** The model’s response partially aligns with the definition but lacks specific details. **Bad (1):** The model’s response does not align at all, indicating a lack of understanding of the concept.”

### Domain knowledge Prompt

**Shot\_Angle\_Recognition:** "Describe the shot angle used in the given video. Respond with only the shot angle used.  
\nChoose from: \naerial: captured from an elevated vantage point\noverhead: when the camera is placed directly above the subject.\neye-level: the camera is positioned directly at the subject's eye level.\nhigh-angle: camera points down on the subject from above.\nlow-angle: camera is positioned below the eye level and looks up at the subject. \nRemember, ONLY respond with one shot angle, do not include any other words."

**Shot\_Type\_Recognition:** "Describe the shot type of the given video. Respond with only the shot type detected. \nChoose from: \nthree-shot: has three characters in the frame.\nover-the-shoulder: shows the main subject from behind the shoulder of another subject\ntwo-shot: features two subjects facing forward in the frame\ngroup-shot: features a group of subjects in the shot\nsingle: captures one subject.\nRemember, ONLY respond with one shot type, do not include any other words."

**Shot\_Motion\_Recognition:** "Describe the shot motion used in the given video. Respond with only the shot motion used.  
\nChoose from: \nhandheld: shot is taken with the camera being supported only by the operator's hands and shoulder.\nzoom: when the camera moves forward and backward adding depth to a scene.\nlocked: shot is taken without shifting the position of the camera.\npan: shot is when the camera is moving horizontally while its base remains in a fixed position.\ntilt: when the camera moves vertically up or down with its base fixated to a certain point.\nRemember, ONLY respond with one shot motion, do not include any other words."

**Shot\_Location\_Recognition:** "Describe the shot location used in the given video. Respond with only the shot location used. \nChoose from: \noutdoor: shot is taken outdoors.\nindoor: shot is taken indoors.\nOnly one answer is possible.\nRemember, ONLY respond with one shot location, do not include any other words."

**Shot\_Subject\_Recognition:** "Describe the shot subject used in the given video. Respond with only the shot subject included. \nChoose from: \nAnimal \nLocation \nObject \nHuman \nLimb \nFace \nText \nOnly one answer is possible.\nRemember, ONLY respond with one shot subject, do not include any other words."

**Cut\_Type\_Recognition:** "Describe the cut type used in the given video. Respond with only the cut type used. \nChoose from: \nreaction-in/reaction-out-cut: A cut to the reaction of a subject (facial expression or single word) to the comments or actions of other actors, or a cut after the reaction\nmatch-cut: Cut from one shot to another by matching a concept, an action or a composition of both\nemphasis/deemphasis: Cut from wide to close within the same shot, or the other way around\nsmash-cut: Abrupt cut from one shot to another for aesthetic, narrative, or emotional purpose\ncut-on-action: Cutting from one shot to another while the subject is still in motion\ncross-cut: Cutting back and forth within location\nCut-away: Cutting into an insert shot of something and then back.\nRemember, ONLY respond with one cut type, do not include any other words."

**Color\_Style\_Recognition:** "Describe the color style used in the given video. Respond with only the color style used. Choose from: \ncold: tones are dominated by blue, cyan, and green hues, giving a cool, calm, or detached feel. \nmiddle: maintain a balanced color representation without leaning towards cool or warm, creating a natural look. \nwarm: tones emphasize red, orange, and yellow, creating a warm, inviting, or nostalgic feel. \nRemember, ONLY respond with one color style, do not include any other words."

**Speed\_Recognition:** "Describe the speed in the given video. Respond with only the video speed. \nChoose from: \nslow: Increases the frame rate during recording, resulting in slowed-down playback that emphasizes details and creates a dramatic effect. \nnormal: Captures and plays back at the standard frame rate \nfast: Reduces the frame rate during recording or speeds up playback, resulting in accelerated movement \nOnly one answer is possible.\nRemember, ONLY response with the shot speed, do not include any other words."

**Shot\_Angle\_Change\_Reason:** "Here are several typical types of shot angles: \naerial: captured from an elevated vantage point\noverhead: when the camera is placed directly above the subject.\neye-level: the camera is positioned directly at the subject's eye level.\nhigh-angle: camera points down on the subject from above.\nlow-angle: camera is positioned below the eye level and looks up at the subject.\n"

**Shot\_Size\_Change\_Reason:** "Provided are several typical types of shot motions:\nhandheld: shot is taken with the camera being supported only by the operator's hands and shoulder.\nzoom: when the camera moves forward and backward adding depth to a scene.\npan: shot is when the camera is moving horizontally while its base remains in a fixed position.\ntilt: when the camera moves vertically up or down with its base fixated to a certain point.\n"

**Cut\_Type\_Reason & Cut\_Type\_Judge:** "Typical cut types involve: \nreaction-in/reaction-out-cut: A cut to the reaction of a subject (facial expression or single word) to the comments or actions of other actors, or a cut after the reaction\nmatch-cut: Cut from one shot to another by matching a concept, an action or a composition of both\nemphasis/deemphasis: Cut from wide to close within the same shot, or the other way around\nsmash-cut: Abrupt cut from one shot to another for aesthetic, narrative, or emotional purpose\ncut-on-action: Cutting from one shot to another while the subject is still in motion\ncross-cut: Cutting back and forth within location. \ncut-away: Cutting into an insert shot of something and then back.\n"

**Shot\_Type\_Reason & Shot\_Type\_Judge:** "Typical shot types include:\nthree-shot: has three characters in the frame.\nover-the-shoulder: shows the main subject from behind the shoulder of another subject\ntwo-shot: features two subjects facing forward in the frame\ngroup-shot: features a group of subjects in the shot\nsingle: captures one subject\n"

Figure C.2. Task prompt we used in the inference stage.

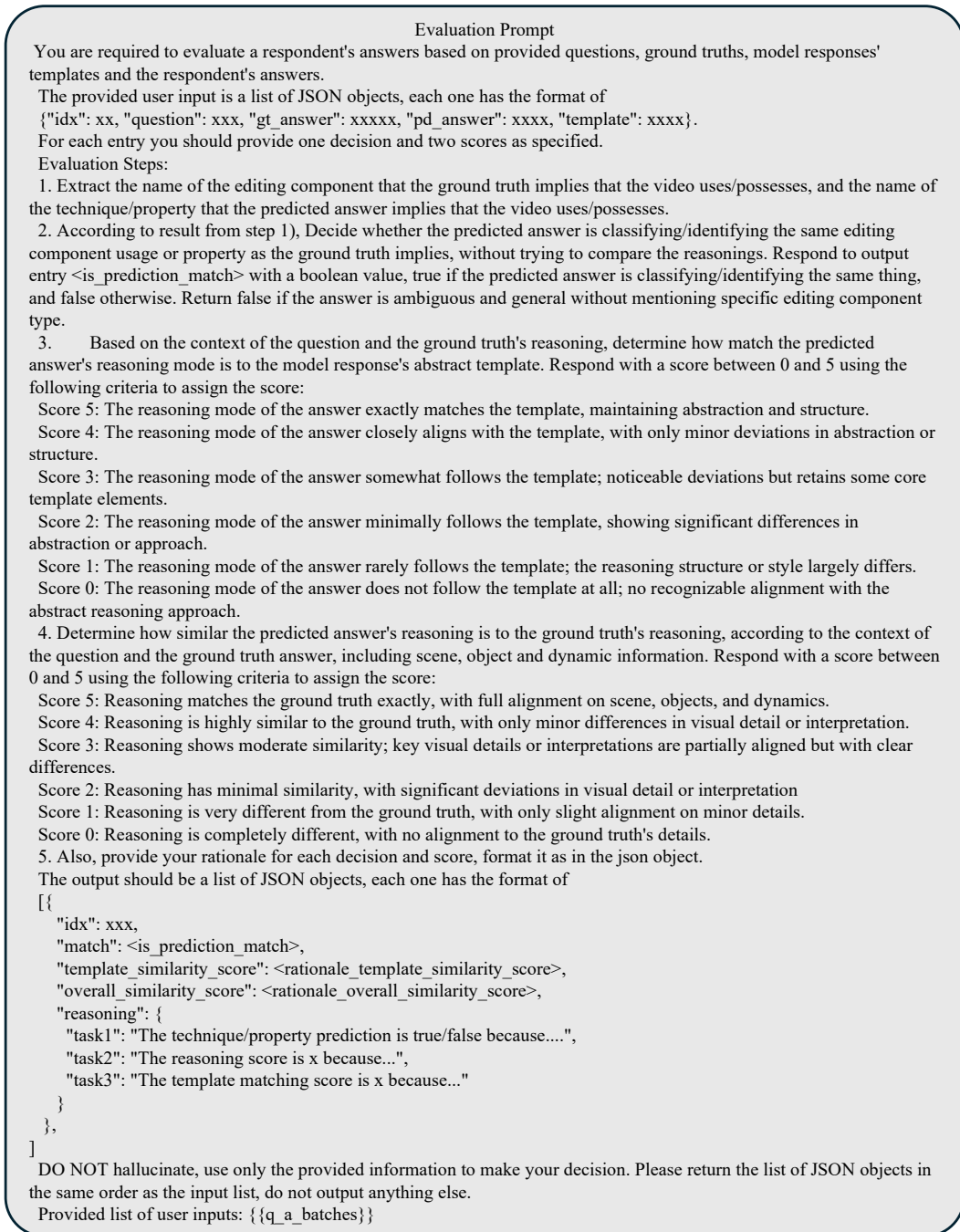


Figure C.3. Evaluation prompt we used in the dataset evaluation stage.

## F. More Result

### F.1. Category-wise analysis

We list category-wise comparison of recognition tasks among Oscars, Qwen2-VL and Gemini Figure F.1. Training on the Oscar dataset enables the model to gain a more balanced and comprehensive understanding of various cat-

egories within each editing dimension. Cases where certain elements in dimensions such as Shot Motion, Shot Angle, and Shot Type have zero recognition accuracy, as observed in Qwen2-VL, are significantly reduced. Compared to Gemini, Oscar also demonstrates more robust comprehension capabilities. For intra-shot and intra-frame element understanding, Oscars excels at distinguishing fine-

Table 7. QA samples of **Concept Experiment** across 10 dimensions

Dimension	Question	Answer
Shot Angle	Can you describe the characteristics of a high-angle shot and how it might influence the viewer’s perception of the subject?	A high-angle shot is when the camera points down on the subject from above, making the subject appear smaller and potentially conveying vulnerability.
Shot location	Can you describe the characteristics of an exterior shot and how it might influence the viewers perception of the setting?	An exterior shot is taken outdoors and typically helps establish the setting or introduce environmental elements crucial to the scene.
Shot Motion	What is a tilt shot, and when would a video editor choose to use this in a scene?	A tilt shot involves moving the camera vertically up or down, often used to emphasize height or follow vertical motion within a scene.
Shot Size	Explain what a medium shot is and how it affects the balance between subject and environment.	A medium shot frames the subject approximately from the waist up, allowing a balance between the subject’s facial expressions and the surrounding environment.
Shot Size	Explain what a medium shot is and how it affects the balance between subject and environment.	A medium shot frames the subject approximately from the waist up, allowing a balance between the subject’s facial expressions and the surrounding environment.
Shot Color	What does shot color mean in the context of video editing?	shot color refers to the overall color tone or palette used in a specific shot, which can convey mood, emotion, and narrative context.
Shot Speed	What does slow motion mean in the context of video editing?	Filming at a higher frame rate to achieve slow motion when played back at a normal rate.
Shot Type	Describe the use of group shots in video editing. What effect do they have on the viewer’s perception of character dynamics?	A group shot features multiple subjects within the frame, emphasizing the relationships or dynamics between characters.
Cut	What is an Emphasis Cut in the context of video editing?	Cut from wide to close within the same shot or the other way around.
Transition	What is a Transition in the context of video editing?	Move from one shot to the next with special visual effects, creating a smooth or intentional visual effect as one clip changes to another.

grained categories, such as Extreme-close-up and Close-up in Shot Size, and Overhead and High-angle in Shot Angle. When recognizing categories that involve both video semantics and visual information synthesis, such as Cross-cut and Emphasis Cut, Oscars shows an improvement of over 30% on average compared to the baseline model Qwen2-VL and performs comparably to the commercial model Gemini. Since the Transition task includes more than 70 categories, it is not included here. Randomly sampled multiple-choice questions effectively reflect the model’s performance in understanding transition types.



Model	Shot Size Recognition				
	EW	W	ECU	M	CU
Qwen2-VL	1	82	32	37	89
Gemini-Pro	26	28	87	82	58
Oscars	92	22	94	72	61

Model	Shot Angle Recognition				
	O	H	L	E	A
Qwen2-VL	63	0	29	82	79
Gemini-Pro	43	61	59	67	0
Oscars	76	20	35	63	68

Model	Shot Type Recognition				
	Th	Tw	S	OTS	G
Qwen2-VL	75	79	76	0	72
Gemini-Pro	79	77	92	36	91
Oscars	85	76	86	47	71

Model	Shot Motion Recognition				
	H	Z	L	T	P
Qwen2-VL	1	1	74	4	61
Gemini-Pro	25	32	74	19	29
Oscars	13	13	67	28	54

Model	Shot Subject Recognition							
	Object	Animal	Cartoon	Location	Limb	Human	Face	Text
Qwen2-VL	5	45	0	85	56	7	94	100
Gemini-Pro	66	59	0	77	58	85	98	100
Oscars	73	5	9	87	68	85	97	100

Model	Shot Location Recognition	
	Indoor Shot	Outdoor Shot
Qwen2-VL	75	79
Gemini-Pro	79	77
Oscars	85	76

Model	Shot Motion Recognition				
	H	Z	L	T	P
Qwen2-VL	1	1	74	4	61
Gemini-Pro	25	32	74	19	29
Oscars	13	13	67	28	54

Model	Cut Type Recognition						
	match-cut	cross-cut	reaction cut	smash-cut	emphasis cut	cut-on-action	cut-away
Qwen2-VL	7	3	100	3	3	0	0
Gemini-Pro	7	3	3	67	33	57	53
Oscars	1	21	73	3	67	7	30

Figure F.1. Category-wise Performance comparison among Oscars, Qwen2-VL [37] and Gemini [34].



## F.2. Details of Performance on general Benchmarks

We present the full set results on Tempcompass and Video-MME(short) in Tab. 8 and Tab. 9. Fine-tuning on VEU data alone enables our model to outperform base model, achieving overall 6.7% improvement on TempCompass benchmark and 2.4% improvement on Video-MME benchmark.

Table 8. Full performance comparison on TempCompass.

Model	action	direction	speed	order	attribute change	Avg.
Qwen2-VL	74.6	39.7	43.4	54.1	53.7	53.2
Oscars	80.8	46.7	47.6	62.6	61.4	59.9

Table 9. Full performance comparison on on VideoMME.

Model	CP	IS	ARec	ORec	ORea	AP	TR	ARea	OCR	SP	TP	SR	Avg.
Qwen2-VL	40.0	82.9	72.3	63.1	72.5	73.0	61.5	68.7	78.9	76.7	83.3	81.5	67.6
Oscars	43.2	86.6	76.6	67.9	68.8	80.3	69.2	65.6	89.5	66.7	77.8	81.5	70.0

## F.3. Evaluation Time Estimation

For benchmark evaluation time estimation, we present the answer generation time cost in Tab. 10, while the answer scoring process is conducted by a single model, taking an average of 25 minutes to complete. Notably, the answer generation time varies depending on model size and efficiency. On average, the evaluation can be completed within 1 hour.

Table 10. Time for Evaluation on a single A100-SXM GPU (min).

Qwen2-VL	MiniCPM-V	GPT4o	Gemini-1.5-pro
45	32	40	33

## G. Qualitative sample of each dimension data

Additional qualitative results are presented in Figure G.1 and Figure G.2. Compared with Qwen2-VL [37] and Gemini [34], Oscars performs admirably across a range of tasks in VEU-Bench. Oscars demonstrate the ability to capture nuanced differences between editing components and reason through video context, showcasing excellent video editing understanding and abstract visual reasoning capabilities.

## H. Limitations and Future Work

Due to the limitations of current mainstream video models in supporting audio understanding, our work focuses solely on video editing comprehension related to visual information, excluding the evaluation of audio-based editing techniques. However, audio plays a critical role in video by contributing to rhythm, narrative, and information delivery.

In future research, we plan to expand the benchmark to include audio, enabling a more comprehensive evaluation of video editing understanding and holistic video comprehension capabilities.

Additionally, as noted in the section on constructing change tasks, current Vid-LLMs struggle to effectively process multiple video inputs. As a result, tasks involving the comprehension and organization of multiple video materials, which are common in editing scenarios, are not included in the current VEU-Bench evaluation. In future iterations, we aim to introduce multi-video editing understanding tasks and anticipate advancements in open-source Vid-LLMs that can better process multiple video inputs and temporal information.

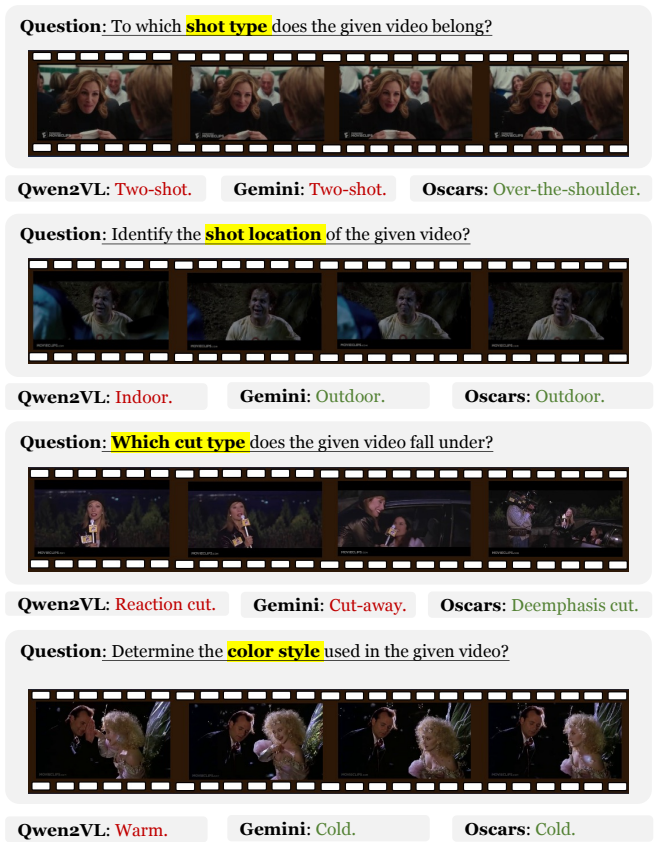
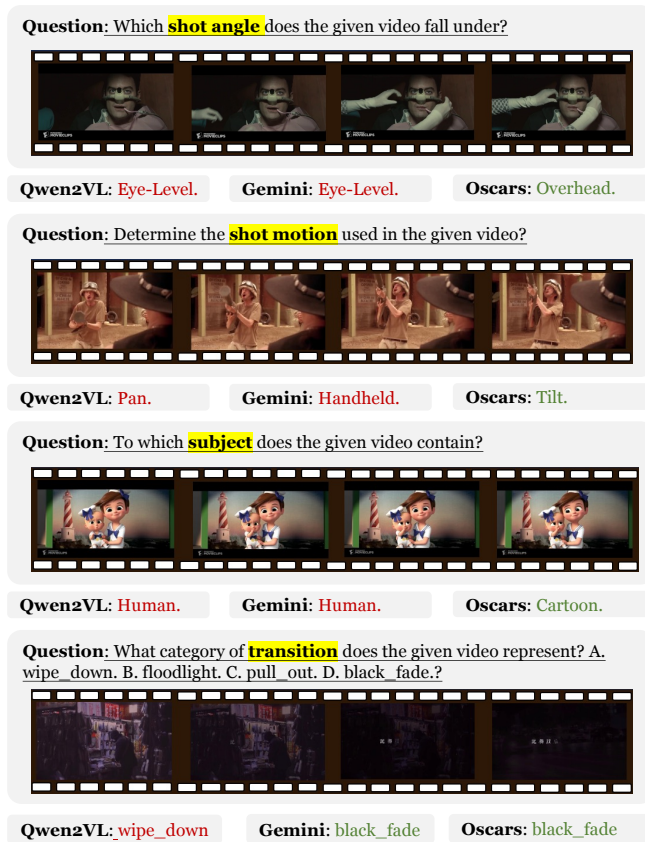


Figure G.1. Qualitative result of Oscars on **recognition task** compared to Qwen2-VL [37] and Gemini [34]

**Question:** In the given video, clarify the **shot angle shift** from one type to another and explain the clue for this change.



**Gemini:** Initially, shot is low angle, looking up at man on boat. This is evident because horizon line is high in the frame, and we're looking up at the boat's structures. The shot transitions to a high angle as the camera looks down on the hook and the men's hands. This shift is rapid, creating a dynamic feel. The sky dominates the frame, and the subjects are seen from above. Then the shot becomes a low angle again. The angle is less extreme than the initial low angle, but it's still looking up at the actors, with the sky as a significant part of the backdrop.

**Qwen2VL:** In the given video, the shot angle shifts from an overhead angle to an eye-level angle. The clue for this change is the transition in the camera's position relative to the subject. Initially, the camera is placed directly above the subject, as indicated by the overhead angle. Then, the camera moves to a position where it is positioned directly at the subject's eye level, as indicated by the eye-level angle. This change in angle provides a different perspective on the subject and can be used to create a more dynamic and engaging visual experience.

**Oscars:** The shot angle of the given video transitions from a high-angle to a low-angle. Initially, the camera looks down at the person fishing beside the boat. As the person descends, the camera angle shifts upward to focus on two men as they lean against the railing.

**Question:** In the given video, identify if the **shot location changes** from indoor to outdoor or from outdoor to indoor, and explain the **reason** for this change.



**Gemini:** The given video has several changes in shot locations from indoors to outdoors. Between 00:00:00 to 00:00:05, the scene is indoors, showing a child on a bed and an adult near him. The room's ambient is dark. From 00:00:06, the setting changes to outdoors, portraying a race car track in a fast motion blur. The shots focus on a man commenting and the car speeding on the track. The backdrop is dark and bright with track lights.

**Qwen2VL:** The shot location changes from indoor to outdoor. The reason for this change is that the scene transitions from a dark room where the boy is sitting to a brightly lit room where the man is driving a car.

**Oscars:** The shot location of the given video change from indoor to outdoor. The scene transitions from a dimly lit cockpit interior to a brightly lit, colorful, and dynamic outdoor racing track.

**Question:** In the given video, identify the **shot size change** from one type to another and **state the basis** for your determination.

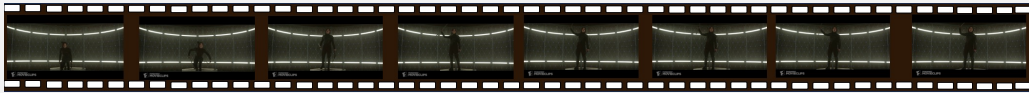


**Gemini:** Extreme close-up to close-up.

**Qwen2VL:** The shot size changes from an extreme close-up to a close-up. The basis for this determination is the transition from showing the entire broken teacup and its fragments to a more detailed view of the teacup's broken pieces.

**Oscars:** The shot size of the given video change from wide to close-up. The camera zooms in on the broken porcelain plate with the words 'Dorothy Porcelain' written on it, shifting the focus from the overall scene of the broken dishes to the details of the plate itself.

**Question:** Identify the **shot motion** of the given video and explain the **basis** for your choice.



**Gemini:** The shot motion used in the provided video is a tilt. Katniss Everdeen, portrayed by Jennifer Lawrence, rises from the platform. The camera moves vertically to follow her. As she stands and speaks, the base of the camera remains fixated in position, and the camera rotates down along a vertical plane following her movements, hence a tilt motion.

**Qwen2VL:** The shot motion in the given video is a pan. The basis for this choice is that the camera is moving horizontally while its base remains in a fixed position, as indicated by the movement of the character across the frame from left to right.

**Oscars:** The shot motion of the given video is zoom. The shot shows Katniss Everdeen in a circular room. The frame size of Katniss changes, appearing larger and closer as the shot progresses, altering the depth perspective of the scene.

**Question:** Identify the **cut type** of the given video and explain the **basis** for your choice.



**Gemini:** The cut type the video clip is a smash cut. In this particular clip, the man is speaking on the phone. The shot then abruptly changes to the next scene, emphasizing that he's been cut off mid-sentence, likely because the other person on the call hung up on him.

**Qwen2VL:** The cut type in the given video is a reaction-in cut. This is because the scene shows a man reacting to something off-screen, and the camera cuts to his reaction. This type of cut is used to emphasize the emotional response of the character to the event or situation being discussed on the phone.

**Oscars:** The cut type of the given video is deemphasis cut. The video transitions from to a closer shot focusing on his face and upper body to a wider shot showing the man's upper body and the surrounding clutter, shifting the audience's focus from the man's expression and reaction to the environment during the phone call.

**Question:** How do you determine the **transition effect** in the given video, and what **clues** reveal it?

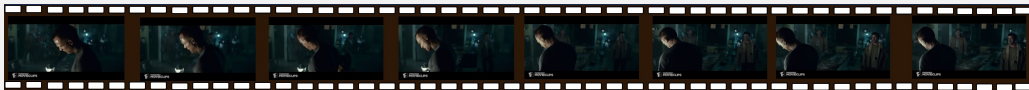


**Gemini:** The transition used in the video is a cube spin transition. It simulates the rotation of a three-dimensional cube, revealing a new image on one of its faces as it turns.

**Qwen2VL:** The transition effect in the video is a wipe effect.

**Oscars:** The transition used in the given video is a cube transition. The first frame, featuring a dog, and the second frame, showing a sunset, are connected through a 3D rotation effect where the content appears mapped onto the rotating faces of a cube.

**Question:** Based on the given video, what kind of **shot type** do you observed in the given video, and what **function** does it serve?



**Gemini:** The shot type in the given video is predominantly a group shot that includes a single shot of the man speaking. The group shot helps establish the scene's location—a dimly lit, seemingly industrial or workshop-like space—and hints at the atmosphere, which feels tense and serious.

**Qwen2VL:** The shot type in the given video is a two-shot, featuring two subjects facing forward in the frame. This type of shot is often used to capture the interaction between two characters, allowing the viewer to see their expressions and reactions to each other. In this case, it serves to highlight the relationship between the two characters and their dynamic.

**Oscars:** The shot type of the given video is three-shot. The three-shot in the video establishes the group dynamics between the three characters. The character in the foreground is looking down, seemingly preoccupied, while the other two characters in the background are observing him, creating a sense of tension and anticipation.

Figure G.2. Qualitative result of Oscars on reasoning and Judging tasks compared to Qwen2-VL [37] and Gemini [34]