VEBENCH: Benchmarking Large Multimodal Models for Real-World Video Editing (Supplementary Materials)

Andong Deng[1,2,‡], Dawei Du[1], Zhenfang Chen[1], Wen Zhong[1], Fan Chen[1], Guang Chen[1], Chia-Wen Kuo[1], Longyin Wen[1], Chen Chen[2], Sijie Zhu[1,*]
[1]ByteDance Intelligent Creation [2]CRCV, University of Central Florida

1. Video Editing Background Knowledge

1.1. Editing Technique Definition

In this section, we define and illustrate the seven fundamental video editing techniques [3] included in **VEBENCH**. Each technique is characterized by its modality (visual or audiovisual) and its unique function in shaping cinematic rhythm, emotion, and continuity.

J-Cut (Audio Leads Video) [Modality: A+V]
The audio from the next scene begins before the visual cut occurs, producing a smooth and anticipatory transition that connects scenes through sound continuity. Named after the shape of the letter "J," this technique allows audio to lead the visual transition.
Movie Example: *The Wolf of Wall Street* (2013), the well-known "Mmm-hmm" conversation between McConaughey and DiCaprio. https://www.youtube.com/watch?v=eyH-a964kAs (0:10)

L-Cut (Audio Follows Video) [Modality: A+V]
The audio from the current scene continues to play after the video cuts to the next one, maintaining auditory continuity and emotional flow. It mirrors the letter "L," where the audio extends beyond the visual boundary.
Movie Example: *The Tree of Life* (2011) — Young Jack opens and closes a door, and the creaking sound continues even after the visual cut, creating a lingering auditory bridge.
https://www.youtube.com/watch?v=eyH-a964kAs (0:53)

Smash Cut (Breaking Continuity) [Modality: V only]
A sudden, jarring cut between two dramatically different scenes in tone, sound, or content. Smash cuts heighten contrast and shock, emphasizing emotional or narrative breaks.
Movie Example: *Reservoir Dogs* (1992) — A sudden alarm blast cuts sharply into the scene as Mr. Pink sprints down the street clutching the stolen diamonds while gunfire erupts behind him, creating a jarring, high-intensity transition.
https://www.youtube.com/watch?v=Q5Y5iqVNrls (0:43–0:49)

Jump Cut (Time Condensation) [Modality: V only]
A discontinuous edit that skips forward in time within the same scene while maintaining spatial coherence. Jump cuts condense time, speed up pacing, or create rhythmic tension in visual storytelling.
Movie Example: *Breathless* (1960), directed by Jean-Luc Godard — The protagonist walks along the street and abruptly "jumps forward" in the same motion path, showcasing one of the earliest iconic uses of the jump cut.
https://www.youtube.com/watch?v=_fcRj0SXYh8 (0:59–1:13)

Cutaway (Narrative Supplementation) [*Modality: V only*]
A cutaway momentarily departs from the main action to show a related shot—such as a reaction, an important detail, or an environmental element—before returning to the primary scene. Cutaways enrich narrative context, emphasize key information, build tension, or provide a smooth visual bridge between edits.
Movie Example: *An American Werewolf in London* (1981) — During the transformation sequence, the film abruptly cuts from the violent werewolf attack to a cheerful cartoon of Mickey Mouse, then cuts back to the attack.
https://www.youtube.com/watch?v=WrIwfImLXOA

Match Cut (Visual Matching) [Modality: V only]
A transition that connects two visually or thematically similar shots to create a symbolic or continuous relationship. Match cuts enhance visual coherence and conceptual association between scenes.
Movie Example: *2001: A Space Odyssey* (1968) — A bone tossed into the air cuts seamlessly to a spacecraft occupying the same position in the frame, creating one of the most iconic match cuts in cinema history.
https://www.youtube.com/watch?v=iGuv1G3-f7w (0:27–0:31)

∗ Corresponding author, sijiezhu@bytedance.com
‡ Work was done during the internship at ByteDance, San Jose, USA

Invisible Cut (Creating Artificial Continuity) [*Modality: V only*]
An invisible cut is a seamless edit that hides the transition between two shots, creating the illusion of a continuous, unbroken take. The cut is typically concealed through fast camera motion, object occlusion, or brief darkness (such as passing behind a wall or a whip pan), preserving immersion and narrative flow.

Movie Example: *1917* (2019) — The film extensively employs invisible cuts to maintain the appearance of a single continuous shot.
https://www.youtube.com/watch?v=ZAQoY3ioci0

1.2. Video Editing Terminology in Interview Video

Interview-style videos commonly rely on two foundational components of professional editing: **A-Roll** and **B-Roll**. Understanding these elements and how they interact are essential for analyzing and reconstructing real-world editing workflows.

A-Roll refers to the primary interview footage, typically featuring the subject speaking directly to the camera or to the interviewer. It carries the core narrative, emotional delivery, and informational content. A-Roll establishes context, provides continuity, and anchors the structure of the video.

B-Roll consists of supplemental visual material intercut with the A-Roll. In interview videos, B-Roll may include movie scenes, archival footage, behind-the-scenes clips, environment shots, reaction shots, or thematic imagery. B-Roll serves multiple editorial purposes: enriching narrative context, illustrating references made by the interviewee, smoothing transitions, reinforcing emotional tone, or providing pacing and visual variety.

Modern cinematic interview formats (e.g., 60 Minutes, Variety, GQ) rely heavily on purposeful B-Roll selection. Editors choose and position these clips with careful intent, to visualize abstract ideas, highlight key roles or achievements, foreshadow upcoming discussion, or maintain viewer engagement through dynamic visual rhythm. In VEBENCH, these elements are essential for constructing the Operation Simulation task. To solve OpSim effectively, a model must understand how A-Roll and B-Roll segments are connected, why specific footage is selected, and how each clip contributes to the broader storytelling logic. This requires reasoning not only about visual-semantic alignment but also about narrative intent, temporal continuity, and the editorial motivations behind placing a particular B-Roll segment at a precise moment in the interview.

2. Annotation Details

2.1. TechRec QA Template

For the **Technique Recognition (TechRec)** task, each question is generated based on the temporal and structural characteristics of the editing technique in the annotated video segment. We design three categories of question templates to reflect different annotation conditions, ensuring coverage of unique, repeated, and long-duration editing events. Below, we summarize the three rule types and provide multiple natural-language templates for each.

Unique Occurrence (Single Cut). If the target editing technique appears exactly once in the video, the question focuses on temporal localization of that single event. Example templates:
- *"When does the* {`CUT_TYPE`} *occur in the video?"*
- *"At what time does the* {`CUT_TYPE`} *happen?"*
- *"Locate the moment when the* {`CUT_TYPE`} *appears in the video."*
- *"During which timestamp does the video apply a* {`CUT_TYPE`}*?"*

requent Cuts (Count-Based). If the technique appears repeatedly (four or more times), we ask the model to count the number of occurrences instead of localizing each one. Example templates:
- *"How many* {`CUT_TYPE`} *occur in this video?"*
- *"Count how many times the video uses a* {`CUT_TYPE`}*."*
- *"What is the total number of* {`CUT_TYPE`} *in the video?"*
- *"How frequently does the video apply the* {`CUT_TYPE`}*?"*

Long-Duration Cut (Transition Range). If the editing technique spans a long temporal range (duration >5s), the question asks for identifying the editing type used over a specific interval. Example templates:
- *"What is the video editing technique used from* {`START`} *to* {`END`}*?"*
- *"Which editing technique is applied during the segment* {`SPAN`}*?"*
- *"Identify the editing technique used throughout the interval* {`SPAN`}*."*
- *"What cut type is present in the time range* {`SPAN`}*?"*

Additional Positional Question Templates. When referring to specific ordinal or positional cuts within the video, we additionally adopt complementary question forms:
- *"What is the video editing technique used at the second cut in the video?"*
- *"What editing technique is applied at the beginning of the video?"*
- *"Which editing technique is used at the final cut of the video?"*

These templates collectively ensure diversity in phrasing

while maintaining structural clarity, enabling robust evaluation of an LMM's ability to recognize and localize video editing techniques across multiple temporal conditions.

2.2. OpSim QA Template

For the **Operation Simulation (OpSim)** task, each question is constructed around the temporal relationship between a reference A-Roll segment and its corresponding B-Roll segment that should be inserted into the editing timeline. The reference A-Roll provides contextual information from the interview video, while the target B-Roll represents the correct movie clip that needs to be selected and temporally localized.

To ensure temporal consistency with real editing workflows, we design two complementary question templates depending on whether the target B-Roll should be inserted *after* or *before* the reference segment. The model must not only choose the correct footage from a candidate list, but also determine the precise timestamps that define the best-fitting segment.

(1) Target B-Roll Appears *After* the Reference Video. When the correct B-Roll follows the reference A-Roll chronologically, the question instructs the model to identify the clip that best continues the current sequence:

> *"What is the best fit to add at the end of the current video clip? Please choose the footage and determine the corresponding timestamps."*

(2) Target B-Roll Appears *Before* the Reference Video. When the appropriate B-Roll precedes the A-Roll in the original edit, the question instead requests the footage that provides the most coherent lead-in:

> *"What is the best fit to add at the beginning of the current video clip? Please choose the footage and determine the corresponding timestamps."*

These two templates allow OpSim to capture realistic editing operations such as narrative setup, context reinforcement, and visual continuation. They also ensure that the temporal intent of each editing action is explicitly communicated to the model, enabling accurate evaluation of both footage selection and temporal localization.

2.3. OpSim Metadata Generation Prompt

The metadata used for the Operation Simulation task consists of a rich, structured JSON representation that captures both global editing attributes and fine-grained temporal scene annotations of an interview video. The prompt for Gemini-2.5-Pro [2] is shown in Figure 1 and Figure 2. As illustrated in the example shown in Figure 3, each video entry contains high-level fields such as `editing_quality`, `interviewee`, and detailed justifications describing the editing style, pacing, and cinematic integration of B-Roll footage. The core of the metadata is the `timestamp_breakdown`, a sequential list of short, non-overlapping segments, each annotated with precise start and end times, a scene description, the inferred source movie (if applicable), and a scene type distinguishing A-Roll interviews, B-Roll cinematic inserts, intro/outro graphics, or narration overlays. For B-Roll segments, we additionally record their narrative purpose, a five-level uniqueness rating, a textual justification explaining why the clip is (ir)replaceable, and a set of five search queries that can be used to retrieve the original source footage. Together, this metadata provides a comprehensive temporal decomposition of the video and serves as the foundation for constructing OpSim questions, selecting reference A-Roll, retrieving candidate B-Rolls, and forming meaningful distractors during dataset generation.

2.4. A/B Roll Pairing Rules

A key component of constructing the Operation Simulation task is the pairing of A-Roll (interview footage) with B-Roll (movie clips, archival footage, or other illustrative visuals). To ensure that the B-Roll associated with each A-Roll segment is semantically meaningful and reflects real editorial intent, we develop a structured annotation protocol that categorizes the narrative function of each A/B Roll pairing into **eight editorial relationship types**. These categories capture how professional editors use B-Roll to support, enrich, or shape the interview narrative.

The eight relationship types are defined as follows:

- **Cause-and-Effect**
  The B-Roll visualizes a direct outcome or consequence of what is being discussed in the A-Roll. It reinforces logical progression or explains how one event leads to another.
- **Illustration**
  The B-Roll provides a literal visual depiction of a concept, object, location, or event referenced verbally in the A-Roll. This is one of the most common uses in interview editing.
- **Contrast / Comparison**
  The B-Roll offers a contrasting example or a comparative visual that frames the current discussion—e.g., showing a past role versus a current role, or contrasting moods and styles.
- **Emotional / Stylistic Reinforcement**
  The B-Roll amplifies the emotional tone, aesthetic, or atmosphere of the interview segment. This type is often used to build mood or emphasize personality traits.
- **Example / Case Study**
  The B-Roll presents a concrete example (e.g., a particular film scene or public event) that supports a general claim or theme mentioned in the A-Roll.

- **Contextual Background**
  The B-Roll provides essential background information—historical, geographical, or biographical—that situates the A-Roll content within a broader context.
- **Flashback / Archival Reference**
  The B-Roll uses past footage, archival material, or retrospective scenes to recall previous events or earlier phases in the interviewee's career.
- **Symbolic / Metaphorical Link**
  The B-Roll serves a symbolic purpose rather than a literal one. It visually expresses a theme, metaphor, or abstract concept mentioned in the A-Roll.

These relationship types are used by annotators when pairing A-Roll and B-Roll segments during dataset creation. By explicitly modeling the editorial intent behind each pairing, VEBENCH captures the narrative structure found in professionally edited interviews. This allows the Operation Simulation task to more faithfully reflect real-world editing workflows, in which selecting the "correct" footage requires understanding not only visual content but also the underlying storytelling logic.

Prompt (P1) Example for Video Metadata Generation

```
You are an expert in film editing and visual media analysis.
I will give you an interview video, and your job is to perform a professional multi-step analysis.
The final output should be in strict JSON format.

---
### Your Tasks

1. **Editing Quality Assessment**
- Evaluate the overall editing quality of the interview video.
- Rate it as one of: "high", "medium", or "low".
- Consider transitions, pacing, timing to music/speech, and use of overlays.
- Provide a detailed justification in "editing_justification".

2. **Dataset Suitability Decision**
- Decide whether this video is suitable for inclusion in a multimodal dataset
  focused on edited cinematic clips.
- Must include meaningful edits, cinematic framing (16:9), and exclude shorts or raw footage.
- Output "Yes"/"No" in "suitable_for_dataset" and explain in "suitability_justification".

3. **Interviewee Identification**
- Identify or infer the interviewee's name from context or transcript.

4. **Timestamp-Based Scene Breakdown**
For each scene, return:
- "timestamp_start": e.g., "0:24"
- "timestamp_end": e.g., "0:31"
- "scene_description": visual content summary
- "source_movie": name or None
- "type": "A-Roll", "B-Roll", "Intro Graphic", etc.
- "b_roll_purpose": narrative function (if B-Roll)
- "broll_uniqueness": "Highly unique" or "Highly replaceable" (if B-Roll)
- "broll_uniqueness_justification": reasoning (if B-Roll)
- "source_search_queries": five YouTube/Google queries (if B-Roll)
```

Figure 1. The prompt (P1) used for video analysis in VEBENCH metadata generation. It asks the model to perform editing-quality evaluation, dataset suitability judgment, interviewee identification, and timestamp-level scene annotation in strict JSON format.

Prompt (P2) Example for Video Metadata Generation

```
---
### Output Format (Strict JSON)

{
  "editing_quality": "high",
  "interviewee": "Cillian Murphy",
  "editing_justification":
  "The editing uses smooth transitions, consistent pacing, cinematic overlay text, and strong color grading.",
  "suitable_for_dataset": "Yes",
  "suitability_justification":
  "The video features widescreen cinematic movie clips integrated into the interview.",
  "timestamp_breakdown": [
    {
      "timestamp_start": "0:00",
      "timestamp_end": "0:04",
      "scene_description": "60 Minutes ticking stopwatch intro graphic.",
      "source_movie": null,
      "type": "Intro Graphic",
      "b_roll_purpose": null,
      "b_roll_uniqueness": null,
      "b_roll_uniqueness_justification": null,
      "source_search_queries": null
    },
    {
      "timestamp_start": "0:24",
      "timestamp_end": "0:31",
      "scene_description": "Cillian Murphy as Oppenheimer walking toward camera in Los Alamos.",
      "source_movie": "Oppenheimer",
      "type": "B-Roll (Movie Clip)",
      "b_roll_purpose": "Illustrates actor's performance in his most recent role.",
      "b_roll_uniqueness": "Unique",
      "b_roll_uniqueness_justification":
      "This clip is directly tied to Murphy's portrayal in Oppenheimer.",
      "source_search_queries": [
        "Cillian Murphy Oppenheimer hat smoking",
        "Oppenheimer walking scene",
        "Cillian Murphy Los Alamos clip",
        "Oppenheimer intro scene",
        "Christopher Nolan Oppenheimer footage"
      ]
    }
  ]
}
```

Figure 2. The prompt (P2) used for video analysis in VEBENCH metadata generation. It asks the model to perform editing-quality evaluation, dataset suitability judgment, interviewee identification, and timestamp-level scene annotation in strict JSON format.

Video Metadata Example


```
"qj5MvD1bpMU": {
  "editing_quality": "high",
  "interviewee": "Timothe Chalamet",
  "editing_justification": "The video exemplifies high-quality broadcast journalism editing. It seamlessly
     integrates traditional A-Roll sit-down interviews, on-location walk-and-talk segments, and a vast array of
      B-Roll from a dozen different films and archival sources. The pacing is excellent, balancing
     introspective moments with dynamic musical performances. Transitions are smooth, and the use of film clips
      is always purposeful, directly illustrating the topics being discussed, such as Chalamet's acting process
     , his specific roles, or his personal history. The audio mix is clean, and the color grading is consistent
      across the interview segments, creating a cohesive and professional final product.",
  "suitable_for_dataset": "Yes",
  "suitability_justification": "This video is highly suitable. It is a professionally edited, widescreen (16:9)
      piece that features a significant amount of cinematic B-Roll from numerous high-profile movies. The edits
      are strong and meaningful, using the film clips to enhance the interview narrative. It contains no
     vertical or user-generated content.",
  "timestamp_breakdown": [
      ...
        {
          "timestamp_start": "0:51",
          "timestamp_end": "0:57",
          "scene_description": "The 60 Minutes stopwatch graphic appears again, indicating a commercial break
    .",
          "source_movie": null,
          "type": "Outro Graphic",
          "b_roll_purpose": null,
          "b_roll_uniqueness": null,
          "b_roll_uniqueness_justification": null,
          "source_search_queries": null
        },
        {
          "timestamp_start": "0:57",
          "timestamp_end": "1:23",
          "scene_description": "A-Roll of Timothe Chalamet being interviewed by Anderson Cooper. The setting is
      a dimly lit space that appears to be a music venue or studio, with a purple-lit background. Chalamet
     discusses his high level of commitment to his roles.",
          "source_movie": null,
          "type": "A-Roll (Interview)",
          "b_roll_purpose": null,
          "b_roll_uniqueness": null,
          "b_roll_uniqueness_justification": null,
          "source_search_queries": null
        },
        {
          "timestamp_start": "1:23",
          "timestamp_end": "1:50",
          "scene_description": "A montage of scenes featuring Timothe Chalamet as Bob Dylan. He is silhouetted
     against a spotlight, then seen from behind as he approaches a microphone on a stage in a large, classic
     theater, singing \"A Hard Rain's a-Gonna Fall\".",
          "source_movie": "A Complete Unknown",
          "type": "B-Roll (Movie Clip)",
          "b_roll_purpose": "To visualize the central topic of the interview: Chalamet's portrayal of Bob Dylan
      and the immense preparation involved.",
          "b_roll_uniqueness": "Highly unique",
          "b_roll_uniqueness_justification": "This is pre-release footage from a major biographical film. It is
      irreplaceable for discussing this specific role and showcases the actor's transformation, which is the
     core of the segment.",
          "source_search_queries": [
            "Timothe Chalamet A Complete Unknown trailer",
            "Timothe Chalamet Bob Dylan singing",
            "A Hard Rain's a-Gonna Fall A Complete Unknown",
            "James Mangold Bob Dylan movie clip",
            "Timothe Chalamet as Bob Dylan on stage"
          ]
        },
        ...
```


Figure 3. Operation simulation metadata example.

3. Open-Source Omni Model Evaluation

To further examine the capability of open-source omni-directional multimodal models on real-world video editing tasks, we evaluate two representative Omni models on VEBENCH, **VITA-1.5** [4] and **VideoLLaMA3** [1]. Both models are tested using the default frame-sampling strategy. Table 1 reports the performance of the evaluated Omni models across all the subtasks. Despite their audio-visual processing capabilities, both models exhibit clear limitations on VEBENCH. They achieve only moderate performance on TechRec and OpSim-FS, and similar to most open-source LMMs, they fail to complete the temporal localization component of OpSim, resulting in near-zero tIoU. These findings further underscore the gap between current open-source Omni models and the demands of realistic video editing workflows, which require nuanced temporal reasoning, multi-video understanding, and fine-grained multimodal perception.

4. OpSim-FS in Frame Index Evaluation Setting

The near-zero scores on the full OpSim reflect a systematic failure in temporal grounding rather than an ill-posed evaluation, which aligns with VEBench's goal of diagnosing the gap between *what* content to use and *how/where* to execute edits. As shown in the Table 2, frame-index (FI) evaluation brings limited improvement overall, which indicates that the difficulty of OpSim stems from intrinsic challenges in temporal decision-making, rather than the evaluation setting.

5. Data Imbalance Analysis

This distribution largely reflects real-world editing practice, where techniques, such as Invisible Cut, are used much less frequently than more common transitions (J-Cut), and often appear only in specific genres or stylistic contexts (e.g., Invisible Cuts in one-shot illusion). We provide a detailed accuracy breakdown by cut type here. As shown in the table, performance remains consistently low not only for rare techniques (Invisible Cut), but also for frequently occurring ones (L-Cut). This indicates that the observed performance degradation reflects the intrinsic difficulty of recognizing subtle editing semantics.

6. More Visualizations

To further illustrate the difficulty and diversity of the **Operation Simulation** task in VEBENCH, we provide two additional qualitative examples. Each visualization compares the predictions of three representative multimodal large models: Gemini-2.5-Pro [2], Qwen3-VL-8B-Instruct [5], and InternVL3-8B [6]. In both examples, the stitched video is composed of five temporally ordered segments: one reference video followed by four candidate option videos (A–D). The model must identify the *only* option containing the most suitable continuation of the reference clip and localize the precise timestamps within that option segment. We show the full stitched timeline, annotated option descriptions, and each model's predicted answer along with the ground truth and temporal IoU score.

In the first example shown in Figure 4, the reference video features Kevin Bacon describing how his character makes exaggerated facial expressions in the film Animal House. The task asks the model to identify which option contains the best segment to append at the end of this reference clip. Ideally, the correct answer should be a moment in Animal House where Kevin Bacon's character is visibly making faces. Option A contains a related scene from the film but does not correspond to the specific moment described. Option B is thematically relevant but still does not show the correct facial-expression sequence. Option C contains a montage of Kevin Bacon's representative works from 1978–2022, which is entirely mismatched for this context. In contrast, Option D includes the iconic "I hate to seem pushy" scene, where Bacon's character performs the exact facial expression referenced in the interview. Gemini-2.5-Pro successfully identifies this segment with a high tIoU, while both Qwen3-VL-8B-Instruct and InternVL3-8B fail completely, selecting unrelated clips.

In the second example, shown in Figure 5, the reference video contains an interview in which Nicholas Hoult explains how he created the zombie vocalization for his role in Warm Bodies. Therefore, when asked to identify the best segment to append at the end of the reference clip, the correct choice should be a scene from the film where Hoult's character produces the distinctive zombie sound he describes. Among the four options, only Option B, the bar scene, includes a moment where the zombie emits a recognizable guttural sound toward another zombie. Gemini-2.5-Pro successfully captures this audio–visual cue and selects the correct segment with a strong tIoU. In contrast, the two open-source models fail entirely, largely due to their lack of audio-processing capability, preventing them from understanding the crucial auditory evidence required for this task.

These examples highlight the unique challenges posed by VEBENCH, including cross-video semantic matching, fine-grained temporal localization, and reasoning about cinematic editing structure.

Table 1. Evaluation results of omni models on **VEBENCH**. "TechRec" denotes the *Technique Recognition* subtask, and "OpSim" denotes the *Operation Simulation* subtask, where "FS" represents *Footage Selection*. The overall score is the average of TechRec and OpSim-FS. **Bold** numbers indicate the best performance.

Models	**Overall (%)**	**TechRec (%)**	**OpSim-FS (%)**	**OpSim (%)**
VITA-1.5 [4]	24.62	22.16	27.08	0.00
VideoLLaMA3 [1]	24.50	23.20	25.80	0.01

Table 2. Operation Simulation performance under frame-index-based evaluation.

Model	Gemini-2.5-Pro	GPT-4o	Qwen3-VL-8B	Qwen3-VL-4B	Qwen2.5-VL-7B	InternVL3-8B
FI tIoU	0.15	0.08	0.04	0.05	0.04	0.00

References

[1] Zesen Cheng Zhiqiang Hu Yuqian Yuan Guanzheng Chen Sicong Leng Yuming Jiang Hang Zhang Xin Li Peng Jin Wenqi Zhang Fan Wang Lidong Bing Deli Zhao Boqiang Zhang, Kehan Li. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 8, 9

[2] Google AI for Developers. Gemini models: Gemini 2.5 pro, 2025. Accessed May 15, 2025. 3, 8, 10

[3] Michael Frierson. *Film and Video Editing Theory*. Routledge, 2018. 1

[4] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 8, 9

[5] Qwen Team. Qwen3-vl. https://github.com/QwenLM/Qwen3-VL, 2025. 8, 10

[6] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 8, 10

Table 3. Accuracy Breakdown of **TechRec** by cut type.

Model	J-Cut	L-Cut	Cutaway	Match	Jump	Invisible	Smash	Overall
InternVL3-8B	24.66	23.08	27.74	31.08	18.87	2.86	21.88	25.21
Gemini-2.5-Pro	38.54	31.24	38.99	35.29	36.36	13.64	29.73	34.65
GPT-4o	10.63	31.65	20.59	22.54	22.64	11.43	15.62	24.68

Model Input:
You are given a stitched video composed of five segments in this exact order:
(1) reference video, (2) option A, (3) option B, (4) option C, (5) option D.
Each segment shows a permanent on-screen tag at the top-left corner: "reference video", "A", "B", "C", or "D".
Use these tags only to identify segments; do not let them bias your choice.

Among the options, there is **ONLY ONE** video that contains one or more segments that can be the best fit to **add at the end** of the reference video. Please choose the video from the options after carefully watching the videos and indicate the supporting time range by selecting the timestamps corresponding to the relevant moments within the option segment and report them on the stitched-video timeline.

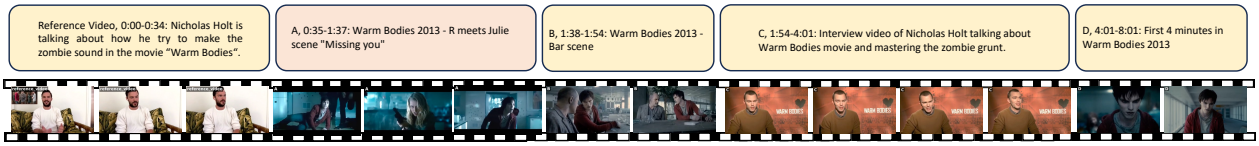
Return your final answer in ONE line only, using the EXACT format below:
<LETTER> <start>-<end>

Gemini-2.5-Pro Response:
D 40:44-40:51 ☑ **tIoU=0.51**

Qwen3-VL-8B-Instruct Response:
C 29.0-31.0 ✗

InternVL3-8B Response:
B 1-15.0 ✗

Figure 4. Qualitative example of Gemini-2.5-Pro [2], Qwen3-VL-8B-Instruct [5], and InternVL3-8B [6].



Model Input:
You are given a stitched video composed of five segments in this exact order:
(1) reference video, (2) option A, (3) option B, (4) option C, (5) option D.
Each segment shows a permanent on-screen tag at the top-left corner: "reference video", "A", "B", "C", or "D".
Use these tags only to identify segments; do not let them bias your choice.

Among the options, there is **ONLY ONE** video that contains one or more segments that can be the best fit to **add at the end** of the reference video. Please choose the video from the options after carefully watching the videos and indicate the supporting time range by selecting the timestamps corresponding to the relevant moments within the option segment and report them on the stitched-video timeline.

Return your final answer in ONE line only, using the EXACT format below:
<LETTER> <start>-<end>

Gemini-2.5-Pro Response:
B 123.0-137.9 ☑ **tIoU=0.37**

Qwen3-VL-8B-Instruct Response:
C 8.0-14.0 ✗

InternVL3-8B Response:
C 1-11.0 ✗

Figure 5. Qualitative example of Gemini-2.5-Pro [2], Qwen3-VL-8B-Instruct [5], and InternVL3-8B [6].