

# VlogReward: Learning Multi-Dimensional Evaluation for Vlog Editing

Yexiang Liu, Wen Zhong, Sijie Zhu, Xin Gu, Fan Chen,  
Junxian Duan, Jie Cao, Longyin Wen, Zhenfang Chen

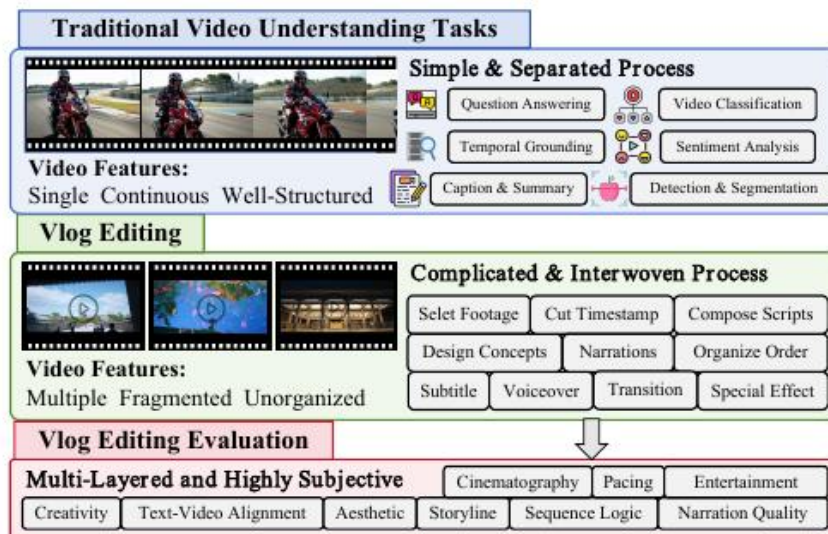
**Key question: can multimodal models evaluate and improve vlog edit plans?**

# Motivation: Why Vlog Editing Needs a Reward Model

## Core challenge

**Vlog quality is not just video understanding; it is an editing judgment.**

- Raw clips are fragmented, redundant, and weakly organized.
- An edit plan combines shot order, cut timestamps, subtitles, voice-over, and pacing.
- The key criteria include both factual alignment and subjective aesthetics.



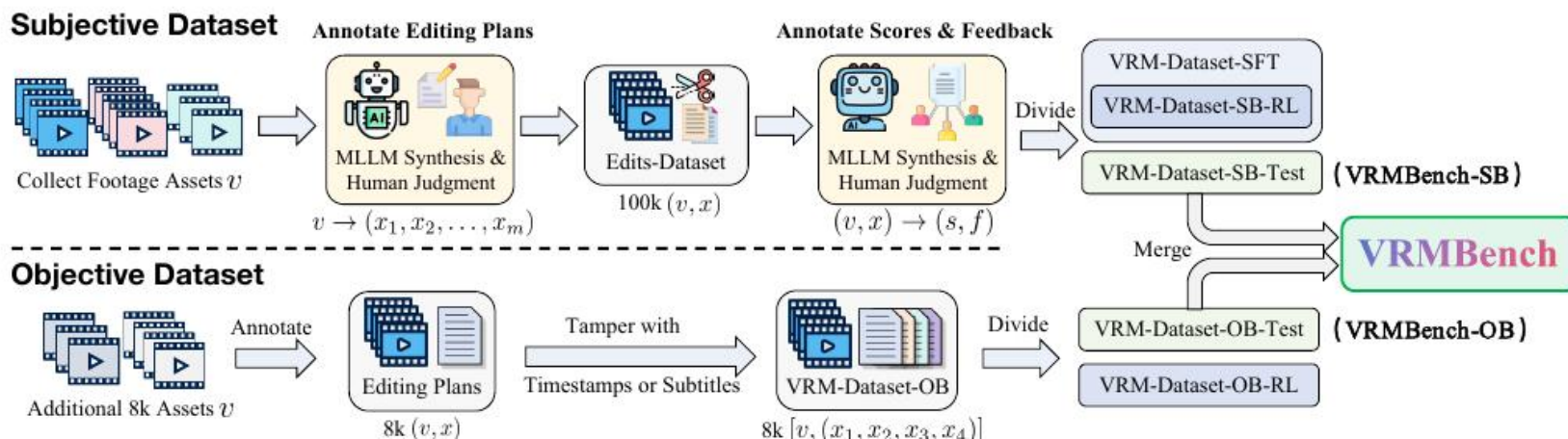
*Vlog editing is more interwoven and subjective than standard video understanding.*

## Six-dimensional evaluation rubric

**Creativity | Consistency | Concept Design | Cinematography | Narration | Pacing**

Goal: produce fine-grained scores and actionable feedback for each candidate edit plan.

# Framework & Data: From Raw Footage to VRMBench



Data collection pipeline: subjective edits, objective tampering, and benchmark construction.

## Reward model output

$$M(v, x) \rightarrow (s, f)$$


- s: six integer scores from 1 to 5
- f: diagnostic feedback per dimension
- Used for selection, RL reward signals, and iterative refinement

## Dataset scale

**100k SFT edits**  
**40k RL edits**  
**400 benchmark groups / 1.6k plans**

VRMBench contains subjective cases and objective factual-error cases.

# Method: SFT + GRPO with Inter-Group Comparison Reward



**Video Assets  $U_n$**

```

{
  "shot_timestamp": "00:07:00:11",
  "final_timestamp": "00:00:00:04",
  "voiceover": "Okay guys, you are NOT going to believe how beautiful Nara is!",
  "title_text": "A Day in NARA",
  "subtitle_text": "Okay guys, you are NOT going to believe how beautiful Nara is!"
},
  .....
  {
    "shot_timestamp": "01:21-01:26",
    "final_timestamp": "00:24-00:29",
    "voiceover": "What an incredible trip. I'm leaving with so many memories.",
    "title_text": "",
    "subtitle_text": "What an incredible trip. I'm leaving with so many memories."
  }
}
        
```

**Vlog Editing Plan  $\mathcal{L}$**

**Creativity: 3**  
The plan outlines a functional but predictable travel vlog. The narrative follows a simple "look at this beautiful place" structure, which is common for the genre but lacks a unique angle or deeper storytelling. The concept of "A Day in Nara" is presented as a straightforward montage rather than a narrative with a compelling arc. For instance, the story is a flat sequence of events: arriving, seeing temples, seeing them at night, and leaving. There's no central conflict, unique discovery, or personal transformation that would elevate it beyond a simple visual diary. It's understandable but fails to be memorable or particularly engaging.

**Consistency: 3**  
There are several instances where the visuals are only partially connected to the script's specific message, creating a noticeable gap. For example, at 00:06:00:10, the voiceover says, "The detail is insane," which implies a close-up shot of architecture or craftsmanship. However, the chosen clip (from raw footage at 01:03:01:07) is a wide shot of people walking into a temple courtyard. A more consistent choice would have been a clip like the low-angle shot of the pagoda roof at 00:33. Similarly, at 00:14:00:18, the voiceover describes finding a "super peaceful spot," but the visual (from 00:39:00:43) is a dynamic follow-shot of people walking up stairs, which conveys movement rather than tranquility. While the main storyline is comprehensible, these disconnects weaken the synergy between script and visuals.

**Concept: 4**  
The overall concept is cohesive and well-chosen. The setting of Nara is authentically portrayed, and the character roles of tourists/vloggers exploring the area are logical and consistent. The plan to show them taking photos (at 00:18:00:20) reinforces their role effectively. The on-screen text design is functional for a standard vlog format. All key assets (locations, characters) are present and used logically to support the "Day in Nara" theme.

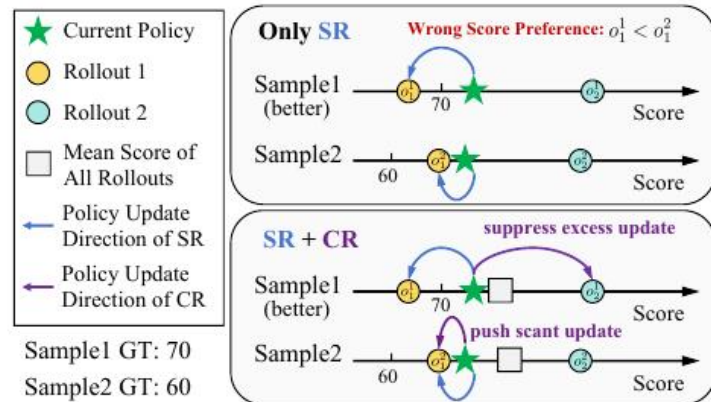
**Pacing: 1**  
The editing plan has a critically flawed and chaotic flow that makes the narrative incomprehensible. The video attempts to show a "Day in Nara," but the shot order is not chronological. It jumps from daytime shots (00:00-00:10) to night shots (00:10-00:18), then illogically cuts back to a daytime shot of the vloggers taking photos (00:18-00:20, using raw clip 00:21-00:23) before moving to sunset shots. This jarring jump from night back to day completely disrupts the story's logic and timeline. This is not an intentional stylistic choice but a fundamental structural error that breaks the narrative sequence and will confuse the viewer.

**Cinematography: 3**  
The raw footage itself demonstrates professional cinematography with stable, well-composed shots. However, the clip selection in the editing plan is functional but suboptimal. The plan often misses the opportunity to use the most impactful part of a clip to match the narration. For instance, to match the voiceover "The detail is insane" (00:06:00:10), the plan selects a generic wide shot (01:03-01:07) instead of a more appropriate detail shot available in the raw footage (e.g., 00:33). This shows a failure to select the clip that best captures the narrative importance described in the script, weakening the final impact.

**Narration: 3**  
The narration is basic but functional. The script relies heavily on superficial statements and common travel vlog clichés like "you are NOT going to believe," "literally obsessed," and "so magical." This is an example of empty content; the narration describes what the viewer can already see without adding personal insight, historical context, or deeper emotional reflection. The wording is often clumsy and fails to create a strong mood or connection with the audience. While it gets the basic point across, it lacks the brilliant writing or emotional delivery needed to elevate the video.

**Multi-Dimensional Scores & Feedback ( $s, f$ )**

Input-output example: raw assets and a JSON plan are mapped to scores and feedback.



Comparison reward adds preference direction across plans.

## Training pipeline

- Cold start with supervised fine-tuning.
- Then train with GRPO on groups of four plans.
- Base model: Qwen2.5-VL-7B-Instruct.

## Total reward

$$r_{\text{total}} = r_{\text{format}} * [(1 - \text{lambda}) * r_{\text{score}} + \text{lambda} * r_{\text{comp}}]$$

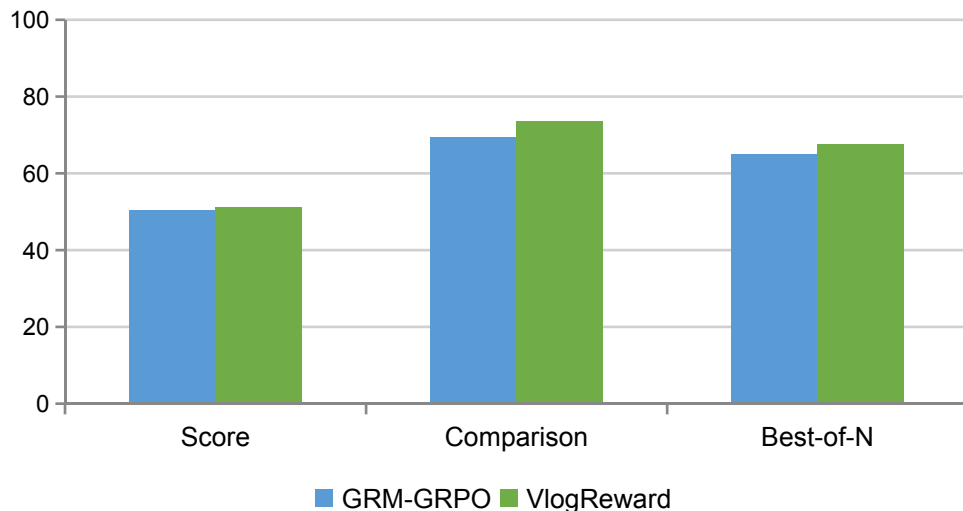
- Score reward aligns each rubric score with the label.
- Comparison reward enforces better plans receiving higher total scores.
- This mitigates direction blindness in score-only GRPO.

# Results: State-of-the-Art Vlog Rewarding



Qualitative example: VlogReward detects chronological disorder while the base model gives identical scores.

Average accuracy on VRMBench (%)



## Key numbers

VlogReward average on VRMBench:  
 Score accuracy: 51.3%  
 Comparison accuracy: 73.5%  
 Best-of-N accuracy: 67.6%

**Effect of comparison reward:**  
**+4.1 points in comparison accuracy**  
**+2.6 points in Best-of-N accuracy**

## Three main takeaways

1. Vlog editing evaluation is a distinct multimodal reward modeling task.
2. A six-dimensional rubric enables interpretable scoring and feedback.
3. Inter-group comparison reward improves preference discrimination.

## Downstream value

Best-of-8 selection improves perceived edit quality in the user study.  
Textual feedback supports iterative refinement of editing plans.  
The comparison reward also transfers to general preference learning.

# Thanks