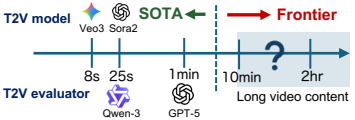
SLVMEval: Synthetic Meta Evaluation Benchmark for Text-to-Long Video Generation

Ryosuke Matsuda[1], Keito Kudo[1], Haruto Yoshida[1], Nobuyuki Shimizu[2], Jun Suzuki[1] ([1]Tohoku University, [2]LY Corporation)

Project Page

CVPR JUNE 3-7, 2026 DENVER COLORADO

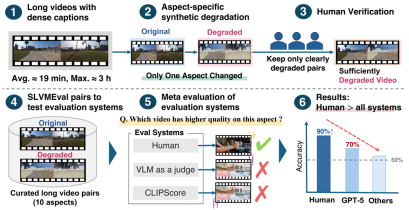
Motivation



As T2V models move toward long-video generation, reliable automatic evaluators are becoming essential for scalable, low-cost feedback. However, it remains unclear whether current evaluation systems can truly understand and judge long videos. To address this gap, we introduce **SLVMEval**, a fine-grained benchmark for meta-evaluating automatic video evaluation systems.

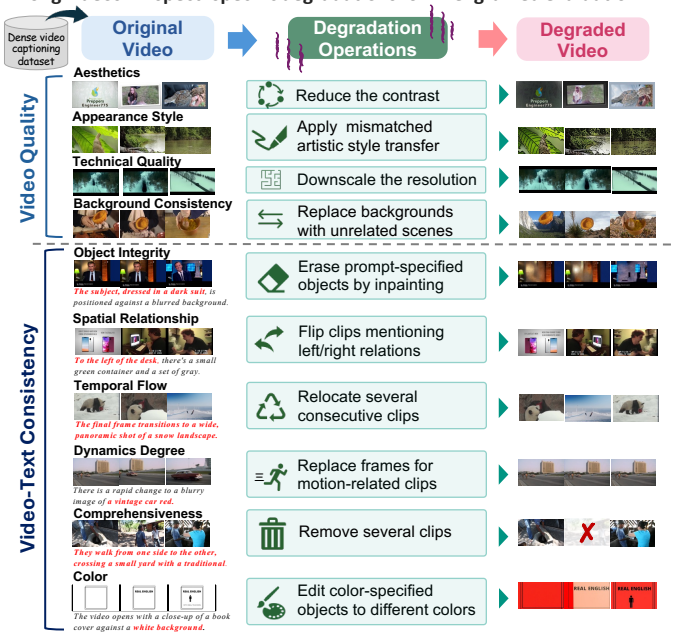
Contribution

- **SLVMEval** enables fine-grained meta-evaluation via controlled degradations across 10 aspects.
- Human-validated benchmark videos **average 19 min**, reaching a approximately **max 3 hrs**.
- Existing evaluators **lag behind humans** in 9 out of 10 aspects, especially for video-text consistency.



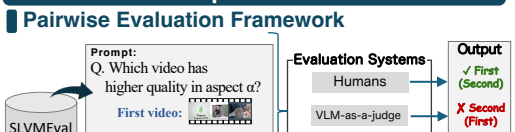
SLVMEval: Dataset Construction

Long videos × Aspect-specific degradations for fine-grained evaluation



Experiments

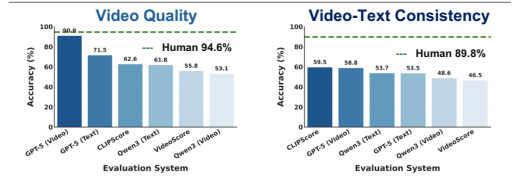
Pairwise Evaluation Framework



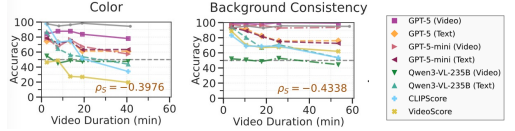
Main Results

Summary: Existing VLMs still struggle with tasks that humans find easy!!

- **Humans** achieve **around 90%** accuracy across all 10 aspects, while **the strongest model** (GPT-5) **falls short of humans** in 9 aspects.
- **Video-Text Consistency** is particularly challenging !!



Sensitivity to Video Duration



- Accuracy generally decreases as video duration increases.
- This tendency appears in most systems and most aspects.