EditDuet: A Multi-Agent System for Video Non-Linear Editing

Marcelo Sandoval-Castañeda [1] Bryan Russell [2] Josef Sivic [2, 3] Gregory Shakhnarovich [1] Fabian Caba Heilbron [2]

[1]Toyota Technological Institute at Chicago [2]Adobe [3]Czech Institute of Informatics, Robotics, and Cybernetics

The Task: A-Roll/B-Roll Editing

Inputs:

A-Roll Video

Media Collection

User Request

```
Slow-paced sequence
of the process of
baking bread, include
hands close-ups [...]
```

Output:

B-Roll Sequence

Previous Works: One-Pass Editing

- **Transcript2Video [5]:** Retrieval-based video editing using a shared text-image feature space.
- **TimelineAssembler [3]:** LLM fine-tuned for generating edited video timelines in EDL format.
- **LAVE [4]:** Human-in-the-loop video editing agent that follows commands in natural language.

The Dataset: EditStock

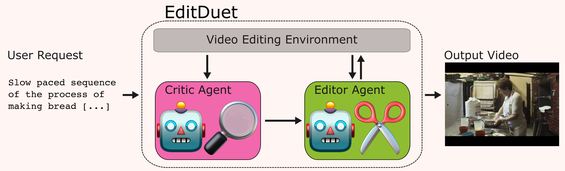
We use real world raw footage obtained through EditStock. In total, this is 1458 minutes of video to be edited down to 21.5 minutes of final cut documentaries.

References

[1] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In CVPR, 2023.
[2] Shikhar Murty, Christopher D. Manning, Peter Shaw, Mandar Joshi, and Kenton Lee. BAGEL: Bootstrapping agents by guiding exploration with language. In ICML, 2024.
[3] Alejandro Pardo, Jui-Hsien Wang, Bernard Ghanem, Josef Sivic, Bryan Russell, and Fabian Caba Heilbron. Generative Timelines for Instructed Visual Assembly. 2024.
[4] Bryan Wang, Yuling Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In ACM IUI, 2024.
[5] Yu Xiong, Fabian Caba Heilbron, and Dahua Lin. Transcript to Video: Efficient Clip Sequencing from Texts. In ACM MM, 2022.

A Two-Agent Iterative System

Our approach consists of two LLM agents: an Editor and a Critic. They interact between themselves and with the environment to satisfy a given user request.



The Environment

Our video editing environment is inspired by real video editing software. It consists of a video collection, a search panel, and a timeline.

The Editor

Our editor agent takes as input the editing environment and a string of natural language feedback from the critic. It has access to tools that are common actions in video editing software, and an action `DONE` to signal that it is satisfied with the current timeline. At the end of its run, the resulting object is a draft timeline.

Tools

```
search_collection        add_to_timeline
remove_from_timeline     switch_clips
move_clip                DONE
```

The Critic

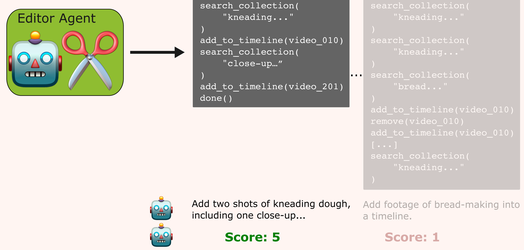
Our critic agent takes as input the editing environment and a user request in natural language. It has access to two tools: `give_feedback` to signal to the editor that there are still some changes to be made or `RENDER` to render the video and end the process.

Tools

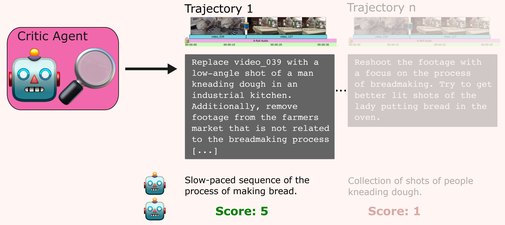
```
give_feedback
RENDER
```

Two-Step Self-Supervised Exploration

The Editor is tasked with exploring different potential sequences of actions starting from a randomly initialized timeline. These sequences are labeled with natural language feedback and scored using external LLMs. High scoring sequences are kept as in-context examples.



The Critic is tasked with improving randomly initialized timelines by interacting with an Editor. Its outputs are labeled with potential user requests and scored using external LLMs. High scoring sequences are kept as in-context examples.



Results: User Study

We compare with Transcript2Video [5], VisProg [1], BAGEL [2], a single Editor agent, and an Editor Critic setup without exploration. We used forced choice to compare two candidate videos from the same project.

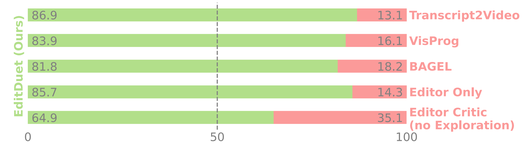
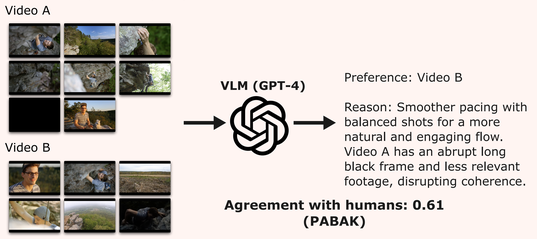


Figure 1. Results of our user study. The green bar indicates preference for our model, the red bar indicates preference for the competing model.

Scaling Up Evaluation

Using a LLM-as-judge approach, we are able to scale up the evaluation. A VLM takes two storyboards as input and is prompted to choose the one that is better edited.



Results: Automatic Evaluation

We extend our user study with results from the automatic evaluation described above. Performances remain consistent, with our model being preferred against all candidate systems.

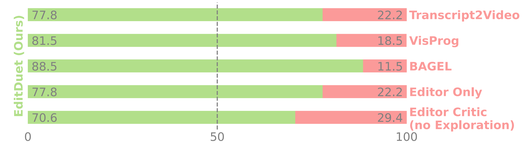


Figure 2. Results with our automatic evaluation.

For qualitative results, see the iPad!