From Long Videos to Engaging Clips: A Human-Inspired Video Editing Framework with Multimodal Narrative Understanding

Xiangfeng Wang[*1], **Xiao Li**[*2], **Yadong Wei**[2], **Xueyu Song**[2], **Yang Song**[2], **Xiaoqiang Xia**[2], **Fangrui Zeng**[2], **Zaiyi Chen**[2], **Liu Liu**[2], **Gu Xu**[2], **Tong Xu**[†1]
[1]University of Science and Technology of China
[2]ByteDance China
xf9462@mail.ustc.edu.cn, tongxu@ustc.edu.cn
lixiao.dlut@bytedance.com
https://huggingface.co/datasets/wonderful9462/DramaAD

Abstract

The rapid growth of online video content, especially on short video platforms, has created a growing demand for efficient video editing techniques that can condense long-form videos into concise and engaging clips. Existing automatic editing methods predominantly rely on textual cues from ASR transcripts and end-to-end segment selection, often neglecting the rich visual context and leading to incoherent outputs. In this paper, we propose a **H**uman-**I**nspired automatic **V**ideo **E**diting framework (HIVE) that leverages multimodal narrative understanding to address these limitations. Our approach incorporates character extraction, dialogue analysis, and narrative summarization through multimodal large language models, enabling a holistic understanding of the video content. To further enhance coherence, we apply scene-level segmentation and decompose the editing process into three subtasks: highlight detection, opening/ending selection, and pruning of irrelevant content. To facilitate research in this area, we introduce DramaAD, a novel benchmark dataset comprising over 2500 short drama episodes and 500 professionally edited advertisement clips. Experimental results demonstrate that our framework consistently outperforms existing baselines across both general and advertisement-oriented editing tasks, significantly narrowing the quality gap between automatic and human-edited videos.

1 Introduction

In recent years, the rapid increase of online video content has significantly enriched users' viewing experiences. However, it has also increased the cognitive burden of content consumption, particularly for long-form videos. Video editing—specifically, condensing lengthy videos with low information density into short, engaging, and information-rich clips—offers a potential solution. Such edited clips not only alleviate users' browsing pressure but also facilitate easier content sharing. With the rise of short video platforms, the demand for efficient video editing has grown substantially. Currently, video editing is largely performed manually by content creators, who must watch the entire video and identify highlights for re-creation. Although manually edited clips tend to be of higher quality, this process is labor-intensive and inefficient, making it unsuitable for large-scale applications. As a result, automatic video editing has emerged as a promising solution to address these challenges.

Early approaches to automatic video editing typically adopt end-to-end frameworks for training editing models. Some studies (Chen et al., 2016; Podlesnyy, 2020; Argaw et al., 2022) formulate the task as a sequence selection problem, effectively performing subtraction on the original video by removing less informative segments. Other works (Hu et al., 2023) treat it as a next-shot retrieval problem, performing addition by selecting and assembling shots to construct the edited video. However, these early methods were limited by the capacity of the models and lacked explicit understanding of narrative structures within videos. Consequently, the generated clips often failed to exhibit coherent storylines or logical progression.

In recent years, the emergence of large language models (LLMs) and multimodal large language models (MLLMs) has significantly advanced the capability for video narrative understanding, making content-aware automatic video editing increasingly feasible. Existing LLM-based editing approaches, such as FunClip [1] and AI YouTube Shorts Generator [2], typically employ automatic speech recognition (ASR) to transcribe spoken dialogue into text, which is then processed by LLMs to infer the

[*]Equal contribution.
[†]Corresponding author.

[1]https://github.com/modelscope/FunClip
[2]https://github.com/SamurAIGPT/AI-Youtube-Shorts-Generator

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 2764–2781
November 4-9, 2025

narrative and select segments for editing.

However, these ASR-based methods primarily rely on dialogue and thus lack access to the rich visual information present in videos. Thus, they often fail to capture non-verbal cues such as facial expressions, gestures, and other visual context, leading to an incomplete understanding of the video content. In addition, we observe that existing methods tend to adopt relatively simplistic editing strategies. They often rely on prompt engineering to elicit end-to-end predictions from the model regarding which segments should be included in the final clip. However, this approach frequently results in abrupt transitions between shots, which can negatively affect the overall viewing experience. In summary, automatic video editing faces two major challenges: (1) an incomplete understanding of video narratives caused by insufficient fusion of visual and textual modalities, and (2) the absence of systematic editing strategies, often leading to abrupt transitions and incoherent storytelling.

To address these challenges, we propose HIVE, an automatic video editing framework based on multimodal narrative understanding. The HIVE considers both the visual content and the underlying storyline, enabling the generation of diverse and engaging edited videos. Specifically, to achieve accurate video comprehension, we incorporate not only character extraction and dialogue analysis but also leverage MLLMs to generate narrative summaries of video content, thereby compensating for missing visual information. To enhance the coherence and viewing experience of the edited clips, we adopt an improved video scene segmentation (Rao et al., 2020), instead of conventional shot detection (Fu et al., 2017; Soucek and Lokoc, 2024), to divide the video into semantically complete scenes. This ensures that clip boundaries do not interrupt ongoing dialogue or character actions.

Furthermore, to achieve human-level editing quality, we conducted a detailed study of the editing techniques commonly employed by professional editors. Inspired by their strategies, we decompose the editing task into three relatively simple sub-tasks: highlight detection, opening/ending selection, and irrelevant content pruning. This decomposition allows the LLM to progressively refine the video in a manner analogous to human editors. Our experimental results show that this step-by-step approach leads to more coherent and polished results compared to end-to-end editing methods.

To support research in automatic video editing, we present DramaAD, a novel benchmark dataset constructed from real-world short drama videos. These short dramas, fast-paced serialized episodes typically lasting 1 to 3 minutes, have recently surged in popularity on short video platforms. In practical applications, companies are increasingly interested in repurposing such content into advertisement-style clips through automatic editing, aiming to reduce manual effort and production costs. DramaAD addresses this emerging need by offering both academic and industrial value: it comprises over 2500 raw drama videos and more than 500 professionally edited advertisement clips as references. To facilitate objective evaluation, we also propose a suite of quantitative metrics covering multiple aspects of editing quality. Experimental results demonstrate that our proposed framework significantly outperforms existing baselines in both ad-specific and general video editing tasks, while closing the performance gap with human editors.

Our main contributions are as follows:

- We introduce a multimodal narrative understanding approach that comprehensively interprets video content, ensuring smooth and engaging edited videos.
- We propose an innovative human-inspired automatic video editing framework that decomposes the editing task into three LLM-friendly sub-tasks, enabling superior editing performance.
- To support research in automatic video editing, we have created a dataset of short dramas in the advertising domain, which includes manually edited videos for reference.

2 Related Work

2.1 Automatic Video Editing

In recent years, automatic video editing has gained increasing attention (Soe, 2021; Ronfard, 2021; Koorathota et al., 2021; Wang et al., 2019). Early methods primarily relied on feature extraction for video editing. Pardo et al. (2021) trained a ranking model using contrastive learning techniques, which can rank and identify plausible cuts among all possible transitions between a given pair of shots. Podlesnyy (2020) designed a data-driven editing model that treats the video editing task as a sequence decision problem, allowing the model to determine whether each shot should be retained. Hu

et al. (2023) utilized a pretrained Vision-Language Model to extract editing-related representations and train an editor model by reinforcement learning. These methods lack an explicit understanding of the video content and suffer from a lack of interpretability. Recently LLM-based editing methods have made progress. Some works (Barua et al., 2025; Wang et al., 2024a) utilize LLMs to assist human editors, while other fully automated approaches, such as FunClip, leverage ASR-transcribed speech to achieve video editing.

Compared to these methods, our approach provides a more comprehensive handling of videos and superior editing strategies.

2.2 Video Understanding

Recently, significant progress has been made in LLM-based video understanding. Methods such as VideoChat2 (Li et al., 2024) and LLaVA-Video (Lin et al., 2023) integrate video foundation models with LLMs. VideoChat2 uses a learnable Q-former inspired by BLIP2 (Li et al., 2023) for aligning visual and textual features, while LLaVA-Video employs a simple linear projection trained jointly on image-video data. Other methods, such as ChatVideo (Wang et al., 2023a) and VideoAgent (Wang et al., 2024c), incorporate LLMs within agent frameworks. Furthermore, general-purpose multimodal models, such as GPT-4o (Hurst et al., 2024) and Gemini (Team et al., 2023), also demonstrate powerful visual understanding abilities applicable to video tasks. In addition, Some studies focus on understanding films through audio description (Han et al., 2023), narration (Zhang et al., 2024), and storytelling (He et al., 2024), requiring finer video understanding and stronger narrative flow. In the short-video domain, datasets such as SkyScript-100M (Tang et al., 2024) provide structured textual formats capturing scenes, emotions, and cinematography instructions, thus supporting AI-based video generation tasks. Similarly, Vript (Yang et al., 2025) creates detailed mappings between general videos and fine-grained screenplay text, enabling more precise video description and understanding.

In contrast to these works, our research focuses specifically on video semantic analysis and leverages multi-dimensional information to achieve deeper video understanding.

3 Method

To tackle automatic video editing, we propose a two-stage framework: (1) Comprehensive content analysis, including characters, dialogues, character-dialogue matching and video scene segmentation, to obtain accurate narratives for each scene. (2) An editing strategy inspired by human editors, which detects highlight clips, selects attention-grabbing openings & suspenseful endings and prunes redundant content to produce the final edited video.

3.1 Multimodal Narrative Understanding

Our framework combines audiovisual models with MLLMs to convert video content into detailed textual representations, simulating human cognition. It features dialogue analysis, character extraction, character-dialogue matching, video scene segmentation and memory module.

Dialogue Analysis. Films and short dramas often feature dense dialogues, with key plot developments driven by character conversations. To capture this, we first apply ASR technology (Radford et al., 2023), though its transcripts are prone to errors from acoustic interference, background music, and character name ambiguities. To address these issues, we use optical character recognition (OCR) (Bautista and Atienza, 2022) to extract subtitles directly from video frames, providing a reliable reference. Leveraging Doubao 1.5 pro [3], we align and correct ASR outputs based on OCR subtitles, significantly improving transcription accuracy. Additionally, we perform speaker diarization (Horiguchi et al., 2020) to distinguish audio by speaker, enabling fine-grained narrative analysis.

Character Extraction. The goal of character extraction is to identify individuals in videos and build a database of their key information. While annotated data or metadata may sometimes be available, it is often missing and must be generated automatically. To address this, we employ face detection (Deng et al., 2019) and clustering: detecting faces in video frames, extracting facial embeddings, and clustering them into distinct identities. Additionally, we enrich character profiles by extracting character's name or identity from drama dialogues. Integrating these information into the character database can facilitate deeper narrative comprehension.

[3] https://seed.bytedance.com/en/special/doubao_1_5_pro

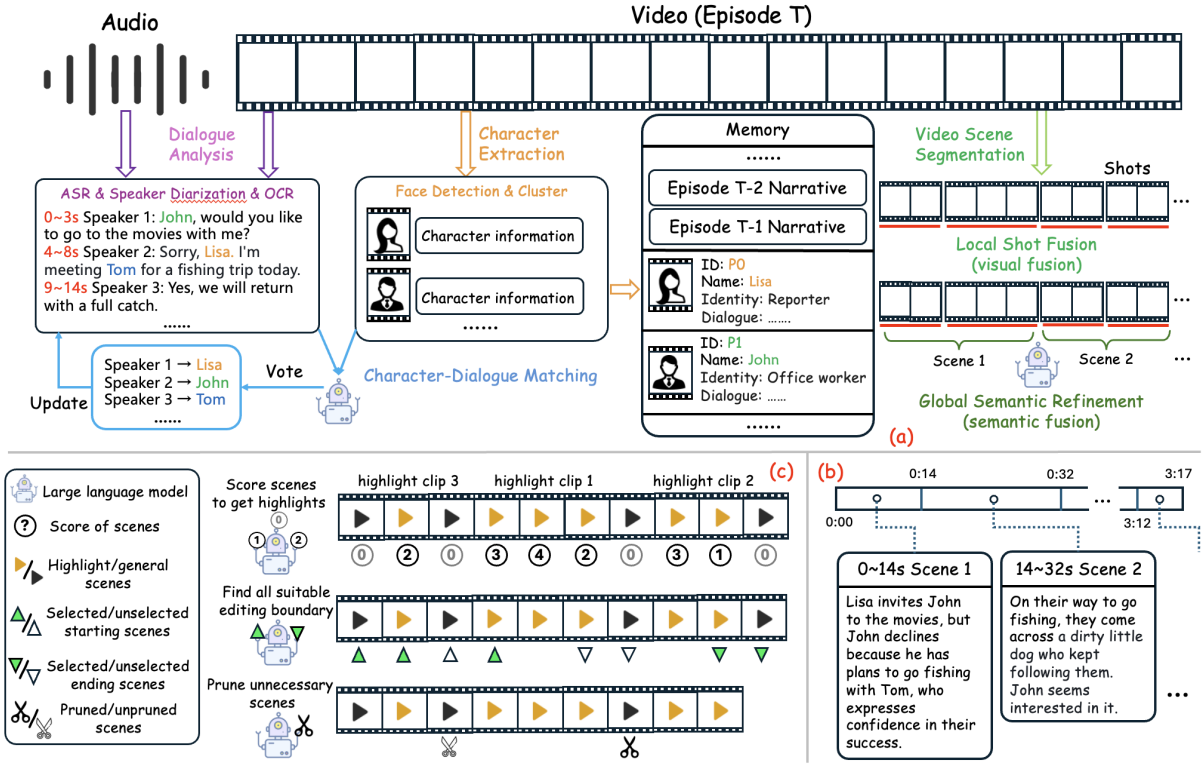

Figure 1: An overview of the proposed framework. The video understanding module **(a)** extracts video information across multiple dimensions, such as dialogues and characters, etc., and segments the video into semantically complete scenes. Consequently, we obtain a sequence of scenes along with their corresponding narrative descriptions **(b)**. This content is then fed into the editing module **(c)**, which identifies highlights, selects appropriate editing boundaries, and removes unnecessary scenes as human editors do (take highlight clip 1 for example).

Character-Dialogue Matching. To enable human-like video understanding, we introduce a specialized module that links identified characters to their dialogues and infers their relationships. After applying speaker diarization, dialogue segments are initially unlabeled. To align the speaker with the spoken content, we input the character's facial image and dialogue information into Doubao 1.5 pro to reason and infer the speaker's identity. To ensure accuracy, we employ multiple inference iterations followed by a voting mechanism.

Video Scene Segmentation. Video scene segmentation is crucial for dividing video content into coherent semantic units. Traditional methods often rely on low-level visual features, leading to fragmented outputs, especially in dramas with frequent shot transitions. To achieve this, we first apply AutoShot (Zhu et al., 2023) to obtain shot-level boundaries and use Qwen2-VL (Wang et al., 2024b) to merge shots with the same site (e.g. restaurant or office). Subsequently, we further utilize Gemini 2.0 flash to merge them belonging to the same event based on contextual information and offer. Meanwhile, a description of the current event is generated. This approach balances segmentation granularity and semantic completeness, ensuring transition fluency.

Memory. The memory module supports processing requirements for long-form videos and multi-episode short dramas by organizing and managing data around two core dimensions: characters and narrative. For characters, it maintains key information, including visual features, interpersonal relations, dialogues, and narrative trajectories across episodes or video segments. For the narrative dimension, it systematically records scene segmentation results and overall storyline progression. Additionally, this module interacts dynamically with other components, supports real-time updates, and enables multi-dimensional data retrieval, effectively accommodating storyline evolution and facilitating downstream analysis.

3.2 Human-Inspired Editing Framework

To explore better editing methods, we consult several professional advertisement editors and distill their strategies into our editing framework, which

Algorithm 1 Highlight-based Automatic Video Editing Algorithm

Initialization: Highlight rule R_h, Opening rule R_o, Ending rule R_e, Pruning rule R_p
Input: Video composed of n scenes $V = (S_1, S_2, ..., S_n)$, LLM L, Top highlights number k
Output: Edited videos V'
$H = (h_1, h_2, ..., h_k) \leftarrow L(R_h, V)$ ▷ Extract highlight clips H using LLM
$H' \leftarrow \text{sort}(H)$ ▷ Sort highlight clips H by scene score in descending order
Initialize candidate editing boundary set $B \leftarrow \{\}$
for each $h' \in H'[1:k]$ **do**
    $O = \{o_1, ..., o_m\} \leftarrow L(R_o, h', V)$ ▷ Select all suitable opening scenes for h', $O \subseteq V$
    $E = \{e_1, ..., e_n\} \leftarrow L(R_e, h', V)$ ▷ Select all suitable ending scenes for h', $E \subseteq V$
    $B \leftarrow B \cup (O \times E)$ ▷ Enumerate all openings-endings pairs by cartesian product
end for
Initialize edited video set $V' \leftarrow \{\}$
for each $(o, e) \in B$ **do**
    $P = \{p_1, ..., p_l\} \leftarrow L(R_p, V[o:e])$ ▷ Get scenes to be pruned within $V[o, e]$, $P \subseteq V$
    $V' \leftarrow V' \cup \text{Cut \& Splice}(V[o:e] \setminus P)$ ▷ Cut and splice remaining scenes
end for
return V'

consists of three components: Highlight Detection, Opening & Ending Selection, and Content Pruning. Algorithm 1 provides a detailed illustration of the workflow of the editing algorithm.

Highlight Detection. Highlight clips are crucial for producing engaging videos (Yao et al., 2016; Xiong et al., 2019; Badamdorj et al., 2022). In our framework, DeepSeek-R1 (Guo et al., 2025) is introduced to identify highlights. Specifically, we abstract common highlight clips in short dramas and films into a set of rules and use R1 to evaluate whether each scene matches these rules (See Appendix A.3 for more details). Each scene is scored based on the number of matched patterns, while unmatched scenes are labeled as "general scenes" with zero scores. After scoring, adjacent non-zero-scored scenes are merged into "highlight clips", with the total score reflecting their importance.

Opening & Ending Selection. After highlight clips are detected, we select appropriate starting and ending scenes to form the video boundaries. For short drama advertisements, the opening aims to attract viewers' attention, while the ending leaves suspense or concludes climax to encourage clicks. Candidate starting scenes include preceding general scenes and the first scene of each highlight clip; ending candidates include following general scenes and the last scene of highlight clips. The GPT-4o evaluates these candidates (details in Appendix A.4), and several opening-ending pairs are sampled to ensure variety.

Content Pruning. Even after boundary selection, some redundant scenes may remain. To streamline the final video, the GPT-4o analyzes each general scene's relevance to the storyline. Less engaging scenes are pruned, ensuring a smooth narrative flow without unnecessary content. The result is an edited video with a captivating introduction, tightly packed highlights, and a suspenseful conclusion.

It is important to emphasize that we only present an editing framework. The framework itself is model-agnostic, allowing any required models to be replaced with more suitable alternatives depending on the practical context. Thus, our framework ensures both flexibility and adaptability across different applications. All prompts used in our framework are listed in Appendix A.

4 DramaAD Dataset

4.1 Data Collection

Firstly, we collect 30 popular Chinese short dramas from the internet. Subsequently, we hire some advertisement editing experts to create advertisement videos for these short dramas as references. Additionally, we annotate the reference videos with the corresponding multimodal narration and editing timestamps to facilitate future research. In this manner, we construct the proposed dataset, named DramaAD.

4.2 Statistics and Comparison

DramaAD contains 3104 videos, of which 2582 are short drama videos and 522 are edited adver-

Dataset	Domain	Videos	Duration	Resolution	Reference Videos
YouCook2 (Zhou et al., 2018)	Cooking	2K	176h	-	-
InternVid (Wang et al., 2023b)	Open	7.1M	760.3Kh	720P	-
Vript (Yang et al., 2025)	Open	-	1.3Kh	720P	-
AVE (Argaw et al., 2022)	Movie	5.5K	207h	720P-1080P	-
SkyScript-100M (Tang et al., 2024)	Drama	6.6K	2Kh	720P	-
DramaAD (Ours)	**Drama**	2.6K	68h	720P-1080P	**522**

Table 1: Comparison between DramaAD and other related datasets.

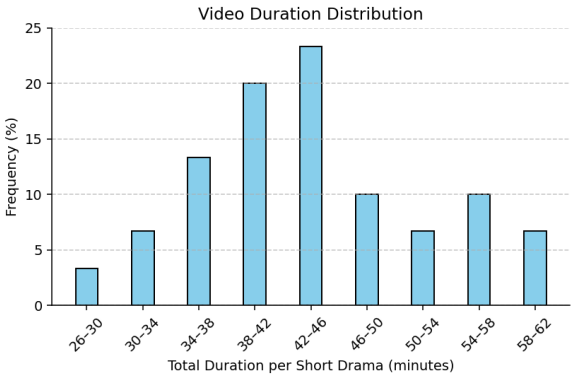

Figure 2: The distribution of the total video duration for each short drama. The average duration is about 40 minutes.

tisement videos. Compared to other datasets, our drama videos have 1080P resolution (the resolution of reference videos ranges from 720P to 1080P) and are in a 9:16 aspect ratio, which aligns better with the trend of increasing popularity of mobile short-form videos. The statistical information of the dataset and comparison with other datasets are presented in Figure 2 and Table 1 seperately.

4.3 Evaluation Metrics

To conduct a comprehensive evaluation, we propose multiple quantitative metrics for assessing both the edited videos and the editing methods. **Diversity.** Diversity is utilized to measure the creativity of the editing methods. Give raw footage, assuming the editing method generates n edited videos $(V_1, V_2, ...V_n)$, the diversity is calculated as follows:

$$\text{Diversity} = 1 - \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} \text{IoU}(V_i, V_j) \quad (1)$$

Smoothness. Abrupt transitions are a common issue in automated video editing. We use a metric called smoothness to evaluate the overall coherence of the edited video. Smoothness is defined as the average uninterrupted duration, meaning the average number of minutes before an unnatural or jarring transition occurs.

$$\text{Smoothness} = \frac{\text{Duration}(V)}{1 + \text{Interruption}(V)} \quad (2)$$

Engagement. Engagement captures the extent to which the edited video maintains viewers' attention. Participants are instructed that they may either switch to 2× playback speed or skip forward using the progress bar if they find any part of the video uninteresting. We define Engagement as the proportion of the video watched at normal speed:

$$\text{Engagement} = \frac{\text{NormalPlayDuration}(V)}{\text{Duration}(V)} \quad (3)$$

Viewing Experience Index. To provide a holistic view of video quality, we define the Viewing Experience Index (VEI) as the product of fluency and engagement:

$$\text{VEI} = \text{Engagement} \times \text{Smoothness} \quad (4)$$

Besides general metrics, we propose two additional metrics specifically for advertising scenarios. **Hook Rate.** It indicates whether viewers will pause to watch the video upon encountering it, reflecting the attractiveness of the video's opening.
Suspense Rate. It measures whether viewers are motivated to continue watching due to the suspense created at the end of the video.

5 Experiments

5.1 Experiment Settings

Data. We sample 220 episodes from DramaAD dataset and select several complete storylines from these videos, which typically span approximately 10 episodes and last 20 to 30 minutes. These videos will serve as the raw footage for the editing process.
Baselines and Reference. We compare our framework with the following baselines.

Method	Strategy	General Scenarios			Advertising Scenarios	
	Diversity	Smoothness	Engagement	VEI	Hook Rate	Suspense Rate
Human (Golden)	0.74	6.84	0.93	6.35	0.87	0.91
End2End (ASR)	<u>0.75</u>	0.65	0.82	0.54	0.64	0.47
End2End (Narration)	0.48	1.28	<u>0.88</u>	1.14	0.62	0.52
HIVE (ours)	0.66	<u>4.48</u>	**0.89**	**4.01**	**0.71**	**0.73**
a. w/o highlight	**0.78**	4.42	0.62	2.74	0.65	<u>0.69</u>
b. w/o boundary	0.54	4.17	0.81	3.38	0.28	0.30
c. w/o pruning	0.66	**5.10**	0.69	<u>3.51</u>	<u>0.69</u>	0.68

Table 2: Evaluation results on DramaAD dataset. The best result among the automatic editing methods is in **bold**, while the second best is <u>underlined</u>.

(1) End2end editing method based on ASR-transcribed dialogue, which is a commonly used editing method. For fairness, we replace the prompts with the same requirements.

(2) End2end editing method based on our video scene segmentation and narratives. Compared to methods in (1), we replace dialogues with scene narratives. For fairness, both methods are provided with the same editing expertise (See Appendix A.6 for more details).

(3) Additionally, human-edited videos in the dataset will serve as the golden editing reference.

Ablation Study Settings. We conduct ablation studies to analyze the contribution of the components in our framework.
a. w/o highlight localization We directly view all scenes as starting and ending candidates.
b. w/o boundary selection We randomly select starting and ending scenes from candidates.
c. w/o pruning All clips from the opening to ending scenes are kept without filtering.

5.2 Experimental Results

Main Results. As shown in Table 2, our proposed method outperforms end-to-end editing approaches in both general and advertisement scenarios.

ASR-based methods yield higher diversity, likely due to richer timestamp information in conversational data, but result in lower fluency, limiting their industrial applicability. Comparing the two end-to-end methods, our video understanding framework improves fluency through semantic-based shot aggregation, albeit at the expense of some diversity.
Ablation Analysis. Ablation studies reveal that removing highlight detection increases diversity but lowers the engagement ratio. The increase in diversity stems from the model's ability to freely select editing ranges across all scenes. Randomly selecting opening and ending scenes significantly weakens advertisement effectiveness and slightly decreases engagement. Omitting scene pruning leads to a lower proportion of engaging content in the final video, but it reduces abrupt visual transitions, thereby improving smoothness. [4]

6 Conclusion

In this work, we present a novel human-inspired framework for automatic video editing, addressing key limitations of existing ASR-based approaches. By integrating multimodal narrative understanding and decomposing the editing process into intuitive sub-tasks, our method achieves superior editing quality with improved interpretability. To further advance research in this area, we introduce DramaAD, the first large-scale short drama dataset with reference advertisement edits, providing valuable resources for both academia and industry. Experimental results demonstrate that our approach significantly narrows the gap between automated and human editing. We hope our work will advance research in the field of automatic video editing.

Limitations

1. The proposed method is currently not compatible with editing techniques such as flashbacks or non-linear storytelling.

2. Since video scene segmentation divides the original video into coarser units compared to shot segmentation, it may reduce the flexibility of editing.

[4] However, since no additional processing is applied to the ending effects when stitching multiple episodes, abrupt transitions may still occur occasionally.

3. Compared to human editing, the proposed method still falls short in terms of metrics such as diversity and smoothness. We hope this work can contribute to advancing research in this area.

Ethics Statement

We strictly follow the ACL Code of Ethics and comply with all guidelines during the research. Our work presents two potential ethical concerns: data source and human editing services.

Data Source. All short drama data in our dataset are authorized by the copyright holders, allowing us to conduct academic research based on these videos.

Human Editing Services. In constructing the dataset, we enlisted experienced editing professionals to generate the reference edited videos. Each video took around 30 minutes to complete, and the experts were compensated fairly based on local rates.

Acknowledgments

This work was supported in part by the grants from National Natural Science Foundation of China (No.62222213, U22B2059).

References

Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. 2022. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *ECCV*.

Taivanbat Badamdorj, Mrigank Rochan, Yang Wang, and Li Cheng. 2022. Contrastive learning for unsupervised video highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14042–14052.

Aadit Barua, Karim Benharrak, Meng Chen, Mina Huh, and Amy Pavel. 2025. Lotus: Creating short videos from long videos with abstractive and extractive summarization. *arXiv preprint arXiv:2502.07096*.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196, Cham. Springer Nature Switzerland.

Jianhui Chen, Hoang M Le, Peter Carr, Yisong Yue, and James J Little. 2016. Learning online smooth predictors for realtime camera planning using recurrent decision trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4688–4696.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C Berg. 2017. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

Yichen He, Yuan Lin, Jianchao Wu, Hanchong Zhang, Yuchen Zhang, and Ruicheng Le. 2024. Storyteller: Improving long video description through global audio-visual character identification. *arXiv preprint arXiv:2411.07076*.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu. 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors. *arXiv preprint arXiv:2005.09921*.

Panwen Hu, Nan Xiao, Feifei Li, Yongquan Chen, and Rui Huang. 2023. A reinforcement learning-based automatic video editing method using pre-trained vision-language model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6441–6450.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Sharath Koorathota, Patrick Adelman, Kelly Cotton, and Paul Sajda. 2021. Editing like humans: a contextual, multimodal framework for automated video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1701–1709.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Alejandro Pardo, Fabian Caba, Juan León Alcázar, Ali K Thabet, and Bernard Ghanem. 2021. Learning to cut by watching movies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6858–6868.

Sergey Podlesnyy. 2020. Towards data-driven automatic video editing. In *Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 1*, pages 361–368. Springer.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10155.

Rémi Ronfard. 2021. Film directing for computer games and animation. In *Computer Graphics Forum*, volume 40, pages 713–730. Wiley Online Library.

Than Htut Soe. 2021. Ai video editing tools. what editors want and how far is ai from delivering? *arXiv preprint arXiv:2109.07809*.

Tomás Soucek and Jakub Lokoc. 2024. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11218–11221.

Jing Tang, Quanlu Jia, Yuqiang Xie, Zeyu Gong, Xiang Wen, Jiayi Zhang, Yalong Guo, Guibin Chen, and Jiangping Yang. 2024. Skyscript-100m: 1,000,000,000 pairs of scripts and shooting scripts for short drama. *arXiv preprint arXiv:2408.09333*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024a. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 699–714.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023a. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, Ariel Shamir, et al. 2019. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.*, 38(6):177–1.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. 2019. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1258–1267.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. 2025. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261.

Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2024. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13657.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wentao Zhu, Yufang Huang, Xiufeng Xie, Wenxian Liu, Jincan Deng, Debing Zhang, Zhangyang Wang, and Ji Liu. 2023. Autoshot: A short video dataset and state-of-the-art shot boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2238–2247.

A Prompt Templates

Here we present our detailed instruction templates which have been translated into English.

A.1 Dialogue Analysis

We instruct LLM to correct ASR-transcribed dialogue with OCR, the prompt template is shown in Table 3.

A.2 Comprehensive Caption

Table 4 shows the details of the scene description we obtained using MLLM as well as multidimensional information.

A.3 Highlight Detection

We defined the highlight scenes of the short dramas according to the gender of their target audience. Table 5 presents the highlight rules for the male audience, while table 6 presents those for the female audience. Table 7 shows the instructions for highlight detection.

A.4 Opening & Ending Selection

Table 8 is about selecting an appropriate opening and ending. Here our prompt is to serve advertising purposes.

A.5 Content Pruning

Please see Table 9 for more details.

A.6 End2End Editing

End2End Editing instruction is shown in Table 10, and the ASR version is presented in Table 11.

Instruction for Correct ASR with OCR

Role
You are an expert in video dialogue content recognition. You're provided with the ASR (Automatic Speech Recognition) results and OCR (Optical Character Recognition) results from the same video, both with associated timestamp information. Your task is to cross-reference these two inputs to correct the ASR results to produce the most accurate speaker identification and dialogue content. The time information in the ASR must remain unchanged.
Input
- ASR: Includes timestamps, speaker identification, and spoken content. Due to potential inaccuracies in detection, both the speaker and the dialogue content might have errors and require correction.
- OCR: Includes timestamps and recognized text from the video's visuals. While OCR text is generally accurate, it may occasionally contain irrelevant visual noise (e.g., text from objects or advertisements), which requires careful filtering.
Output
- Corrected ASR: Includes timestamps, corrected speaker identification, and dialogue content. Time remains unchanged from the original ASR. Corrections are made using OCR where inaccuracies in speaker or content are evident, while irrelevant OCR text is excluded. Speaker labels are adjusted only for clear mistakes to ensure accuracy.
Requirements
1. Cross-reference the timestamp in OCR with the timestamp in ASR to adjust the ASR speaker and spoken dialogue content.
2. Correct typical errors such as phonetic spelling issues in ASR by using OCR text with matching timestamps.
3. Ensure that the corrected ASR speaker and content align logically within the context of the video.
4. Maintain the original timestamp information from the ASR, regardless of corrections made.
5. Filter out irrelevant OCR content such as noise, and avoid overwriting ASR with unrelated visual information.
ASR:
{ASR}
OCR:
{OCR}

Table 3: Instruction template for Correct ASR with OCR

Instruction for Comprehensive Caption

Role
You are a professional film and drama captioning expert. Your task is to produce accurate and coherent narrative descriptions of key scenes in a video segment by integrating multimodal inputs while maintaining logical plot continuity.
Input
- Current Video Segment: The specific video clip or sampled frames corresponding to the segment that needs to be summarized.
- Character Information: A list of characters appearing in this segment and their roles in the drama.
- Dialogue: A transcript of all dialogues spoken within this scene, including timestamps and speaker attributions.
- Previous Segment Context: A concise overview of important events or character dynamics from the previous segment.
Output
- Comprehensive Description:
A coherent and context-aware summary of the segment that integrates characters, dialogue, and contextual information to reflect its significance within the broader narrative. Key events, character interactions, and logical plot connections should be naturally framed, avoiding verbatim dialogue and focusing on emotional and story progression.
Requirements
1. Holistic Scene Summarization:Generate a coherent narrative that captures key events, character motivations, relationships, and emotional flow while emphasizing the scene's role in advancing the plot or developing character arcs.
2. Continuity Maintenance: Ensure the summary remains consistent with the Previous Episode Summary and Current Episode Summary, connecting past interactions or unresolved conflicts to the current scene within the broader storyline.
3. Incorporate Dialogue and Key Interactions: Integrate key dialogues naturally into the description to highlight important developments, reframing them fluently into the narrative without listing them verbatim.
4. Context Sensitivity: Use all provided inputs (Character Information, Dialogue, Previous Segment Context and Current Video Segment) to create a summary aligned with both scene details and the overall plot progression.
Current Video Segment:
{Current Video Segment}
Character Information:
{Character Information}
Dialogue:
{Dialogue}
Previous Segment Context:
{Previous Segment Context}

Table 4: Instruction template for Comprehensive Caption

Highlight Definition and Scoring Rules for Male Audience

Category 1 – Modern Male-Lead Stories
This category usually focuses on a male protagonist who hides his true identity or abilities, playing the underdog while secretly being powerful.
Common plots include:
- Unknown Hero Rises: A seemingly ordinary man suddenly enters the scene, shaking up the world around him.
- Return and Revenge: The male lead is forced to stay away for some reason. When he returns, he finds his wife and children have suffered, prompting him to take revenge.
Typical Highlight Moments:
- First Encounter with Female Lead: Their first meeting usually involves intense conflict and serves as an early mini-climax. (3 points)
- Mocked or Doubted: A classic setup where the hero is looked down upon, making the audience eager to see how he proves everyone wrong. (3 points)
- Praised by Powerful Figures: When villains attack the hero, someone with a strong background—like a senior or powerful ally—steps in to support him, subtly hinting at his true status. (3 points)
- Hero Saves the Beauty: A timeless scene—rescuing the female lead when she's in trouble, avenging his wife, or helping her through a career crisis. (3 points)
- Beautiful Woman Appears Out of Nowhere: Whatever the hero wants, it conveniently shows up. For example, an old man whose life the hero saved offers his daughter in marriage, or after the ex-wife scorns him, new admirers suddenly appear. (3 points)
If none of the above apply, you can choose:
- Betting Scenes: Moments involving wagers or dares. (2 points)
- Running into an Ex: Meeting an ex-girlfriend, wife, etc. (2 points)
- Large Crowds: Scenes like auctions, competitions, or any event involving many extras. (2 points)
Category 2 – Historical Male-Lead Stories
These stories often feature a male protagonist who uses memories from a past life or modern-day knowledge to continuously rise in status. They can be split into time-travel or reincarnation subgenres.
Typical Highlight Moments:
- First Pot of Gold: Usually the first big moment in the series—such as using modern knowledge to earn money. (3 points)
- Mocked, Then Fights Back: Again, a classic "looked down upon but rises up" setup. (3 points)
- First Meeting with Female Lead: This encounter usually comes with dramatic tension. (3 points)
Category 3 – General Highlights
These highlight moments apply broadly across all story types:
- Major Climax or Turning Point: This could involve a key character's death, a huge secret revealed, or any major plot shift. (2 points)
- Emotional Resonance: Moments that trigger strong emotions—whether sadness, joy, shock, or fear—that stay with the audience. (1 point)
- Cultural Impact: Scenes that spark social discussion or shift public attitudes. (1 point)
- Cliffhanger: Endings that leave viewers eagerly awaiting the next episode. (2 points)
- Key Plot Twist: Includes major developments like the heroine's identity being revealed or the story's main conflict emerging. (2 points)

Table 5: Highlight scene definition and scoring rules for the male audience.

Highlight Definition and Scoring Rules for Female Audience

Category 1 – Modern Female-Lead Stories These stories are told mainly from the female protagonist's perspective. They typically fall into two types:
- Sweet Romance: Often follows a "married first, fall in love later" plot, with plenty of sweet, heartwarming moments sprinkled throughout.
- Revenge & Glow-Up: The heroine is divorced by her ex-husband, only to thrive and shine even brighter afterward.

Common Highlight Moments:
- Unexpected Accidents: For example, the heroine gets into a car accident. (3 points)
- Arguments & Face-Offs: Such as the heroine confronting a rival female character. (3 points)
- Divorce Scenes: Like the heroine asking the male lead for a divorce. (3 points)
- Identity Reveals or Secrets Uncovered: For instance, it turns out the heroine is a highly skilled expert. (2 points)
- Humor & Comedy: Scenes with funny twists, e.g., the heroine making the CEO male lead ride an electric scooter. (2 points)
- Violence & Intensity: High-drama moments like slapping scenes. (3 points)
- Flirty & Ambiguous Moments: Light teasing or suggestive scenes that keep viewers hooked. (3 points)
- First Encounter Between Leads: Usually the first mini-climax of the show. (3 points)
- One-Night Stand/Pregnancy Plotlines: A one-night stand leads to future encounters filled with potential drama. (3 points)
- Heroine Actively Pursues the Male Lead: "Chasing love" moments are often the sweetest parts. (3 points)
- The heroine in Trouble, Rescued by Male Lead: Typically involves the heroine being mocked, drugged, attacked, or assassinated—only for the male lead to appear heroically. (3 points)
- Heroine's Suffering Moments: Classic melodrama setup, where she is initially oppressed before rising up. (3 points)
- Heroine's Counterattack: She fights back against vicious rivals, ex-husbands, or antagonists. (3 points)

If none of the above apply, you can choose:
- Sweet Interactions Between Leads: Romantic or affectionate scenes. (2 points)
- Male Lead Protecting or Supporting the Heroine: Either defending her or helping with her career. (2 points)
- Pregnancy Reveal: When either lead learns about the pregnancy. (2 points)
- Heroine Pressured to Divorce: Scenes where the ex-husband or in-laws force her to divorce. (2 points)

Category 2 – Historical Female-Lead Stories
These stories usually feature a heroine who uses memories from a past life or modern knowledge to outsmart rivals and win true love.

Common Highlight Moments:
- Pre/Post Time-Travel or Rebirth Scenes: The opening explains the setup, drawing viewers in. (3 points)
- Heroine's Counterattack: She strikes back at jealous female rivals, villains, or underlings. (3 points)
- Heroine in Peril: Trapped or framed by villains, followed by a rescue. (3 points)

If none of the above apply, you can choose:
- Sweet Interactions Between Leads: Hugging, cheek touching, or other intimate moments. (2 points)
- Villain's First Appearance: Introduction of key antagonists. (2 points)

Category 3 – General Highlights
These highlight moments apply broadly across all story types:
- Major Climax or Turning Point: This could involve a key character's death, a huge secret revealed, or any major plot shift. (2 points)
- Emotional Resonance: Moments that trigger strong emotions—whether sadness, joy, shock, or fear—that stay with the audience. (1 point)
- Cultural Impact: Scenes that spark social discussion or shift public attitudes. (1 point)
- Cliffhanger: Endings that leave viewers eagerly awaiting the next episode. (2 points)
- Key Plot Twist: Includes major developments like the heroine's identity being revealed or the story's main conflict emerging. (2 points)

Table 6: Highlight scene definition for the female audience.

Instruction for Highlight Detection

Role
You are a professional short drama editor with a deep understanding of plot development and audience engagement. Your task is to evaluate the highlight level of each scene based on specific scoring guidelines and return the results accordingly.
Input
- Drama Title: The name of the short drama.
- Target Audience Gender: The primary gender demographic of the audience.
- Scene Fragments: Several scene descriptions from a specific episode.

Output
You should output the score for each scene in sequence, formatted as a JSON list, wrapped with <result> tags. The format should look like this:
<result>
[
{"episode": 1, "scene_id": 1, "reason": "Your justification for the score of this scene.", "score": 3},
{"episode": 1, "scene_id": 2, "reason": "Your justification for the score of this scene.", "score": 0},
...
]
</result>
Explanation:
- episode: The episode number.
- scene_id: The scene number within that episode.
- score: The score assigned to the scene.
- reason: The rationale behind the score.

Please note: The input scene fragments may not start from Episode 1. Make sure to keep the episode and scene numbers consistent in your output.
Requirements
1. I will provide you with editing insights that summarize key elements of popular and engaging plotlines. Based on these guidelines, please evaluate each scene as follows:
2. If a scene matches one or more of the provided criteria, add up the corresponding points.
3. If the scene does not meet any criteria or feels plain, assign a score of 0.

{Highlight definition and scoring rules (dependent on audience gender)}

{Scene narration of episode 1 $\sim$ T}

Table 7: Instruction template for highlight detection.

Instruction for Opening & Ending Selection

Role
You are a professional short drama editor with a deep understanding of narrative structure and audience engagement. Your goal is to create captivating edited videos by highlighting the most exciting scenes. For simplicity, the editing strategy is as follows:
- First, select the pre-identified highlight scenes as the core selling points.
- Then, choose suitable starting and ending scenes to wrap around these highlights.
In this task, the highlight scenes are already marked. You need to analyze the drama and select the best start and end scenes accordingly.
Input
- Drama Title: The name of the short drama.
- Target Audience Gender: The primary gender demographic of the audience.
- Drama Plot: The storyline of the drama, is composed of multiple scenes. Each scene contains an episode ID, a scene ID, and a description of its visual content. Some scenes are labeled as follows:
- <Highlight>: Key exciting scenes that serve as the main attraction in the edited video.
- <Optional Start>: Scenes pre-selected as potential starting points.
- <Optional End>: Scenes pre-selected as potential ending points.
Some scenes may carry multiple tags at the same time.
Output
For each scene labeled as <Optional Start> or <Optional End>, you need to decide whether it is suitable to be used as the start or end of the video.
Output your decision as a JSON list wrapped in <result> tags, like this:
<result>
[
{"episode": 1, "scene_id": 1, "thought": "Your reason about if this scene is a suitable start or end scene.", "starting": true, "ending": false},
{"episode": 1, "scene_id": 2, "thought": "Your reason about if this scene is a suitable start or end scene.", "starting": false, "ending": false},
{"episode": 2, "scene_id": 4, "thought": "Your reason about if this scene is a suitable start or end scene.", "starting": false, "ending": true},
...
]
</result>
Explanation:
- episode: episode number.
- scene_id: scene number within that episode.
- thought: your reasoning and decision for each scene.
- starting: true if suitable as the start scene; otherwise false.
- ending: true if suitable as the end scene; otherwise false.
Requirements
For Selecting Start Scenes:
You should analyze the plot and consider:
- Audience Engagement: The opening scene should immediately grab attention and draw viewers in.
- Clarity: Avoid starting with scenes that rely heavily on prior plot context, such as ones mid-event, which might confuse new viewers.
- Introduction Scenes: Scenes that introduce characters, setting, or plot premise can be good starting points.
For Selecting End Scenes:
You should analyze the plot and consider:
- Relevance: Avoid ending on scenes unrelated to the highlight, as they may transition to a new, less exciting storyline.
- Neutral Endings: If a scene doesn't strongly affect the viewing experience, either way, it can still be chosen as an ending.
- Suspense: Prefer ending on scenes that leave a sense of suspense, encouraging viewers to click to find out what happens next.
- Complete Story Arc: It is also acceptable to end where a plot arc is fully wrapped up, giving a satisfying sense of closure while showcasing the highlight.

{Scene narration of episode 1 $\sim$ T with tags}

Table 8: Instruction template for Opening & Ending Selection.

Instruction for Content Pruning

Role
You are a professional short drama editor with a deep understanding of storyline structure. Your goal is to craft engaging and cohesive edited videos. For this task, a relatively attractive video segment has already been pre-selected for you. Your job is to analyze the storyline and remove redundant scenes to make the plot more concise and gripping.
Input
- Drama Title: The name of the short drama.
- Target Audience Gender: The primary gender demographic of the audience.
- Drama Plot: The storyline is composed of multiple scenes. Each scene includes an episode ID, a scene ID, and a description of the visual content.
Additionally, each scene is labeled as either <Highlight Scene> or <General Scene>:
- <Highlight Scene> refers to key moments designed to attract viewers and evoke strong emotions. These are the main selling points of the edited video.
- <General Scene> refers to relatively plain or transitional scenes, which are candidates for possible removal.
Output
For every <General Scene>, you need to decide whether it can be deleted. Generate your decision as a JSON list wrapped in <result> tags, like this:
<result>
[
{"episode": 1, "scene_id": 1, "thought": "Your reasoning on whether this scene should be kept or removed.", "delete": false},
{"episode": 1, "scene_id": 3, "thought": "Your reasoning on whether this scene should be kept or removed.", "delete": true},
{"episode": 2, "scene_id": 2, "thought": "Your reasoning on whether this scene should be kept or removed.", "delete": false},
...
]
</result>
Explanation:
- episode: Episode number.
- scene_id: Scene number within that episode.
- thought: Your reasoning about whether this scene can be removed.
- delete: true if the scene should be removed; false if it should be kept.
Requirements
1. Only remove scenes labeled as <Regular Scene>. Scenes marked as <Highlight Scene> must never be deleted.
2. Do not remove the first or last scene, even if labeled as <Regular Scene>.
3. If there are no deletable scenes, simply output an empty list inside the <result> tag.
4. Maintain the original episode and scene_id numbers as provided.
5. Be cautious when removing scenes. Unnecessary deletions may affect the coherence of the storyline, negatively impacting the viewing experience.

{Scene narration of episode 1 $\sim$ T with tags}

Table 9: Instruction template for Content Pruning.

Instruction for End2End Editing

Role
You are a professional short drama editor with a deep understanding of plot development and audience engagement.
Input
- Drama Title: The name of the short drama.
- Target Audience Gender: The primary gender demographic of the audience.
- Scene Fragments: Several scene descriptions from a specific episode.
Requirements
1. You need to fully understand the content of the input short drama, analyzing the plot and characters to grasp the overall storyline.
2. Based on the scene descriptions, you should edit the drama by preserving the key plot points while ensuring the final cut remains coherent and smooth.
3. Pay special attention to retaining highlight moments within the scenes, as these can significantly enhance the appeal of the edited video. However, always prioritize the overall narrative flow when selecting which highlight moments to include.
I will provide highlight definitions and scoring rules for you.

{Highlight definition and scoring rules (dependent on audience gender)}

Besides highlights, good opening and ending scenes are also crucial.
For Selecting Start Scenes:
You should analyze the plot and consider:
- Audience Engagement: The opening scene should immediately grab attention and draw viewers in.
- Clarity: Avoid starting with scenes that rely heavily on prior plot context, such as ones mid-event, which might confuse new viewers.
- Introduction Scenes: Scenes that introduce characters, setting, or plot premise can be good starting points.
For Selecting End Scenes:
You should analyze the plot and consider:
- Relevance: Avoid ending on scenes unrelated to the highlight, as they may transition to a new, less exciting storyline.
- Neutral Endings: If a scene doesn't strongly affect the viewing experience, either way, it can still be chosen as an ending.
- Suspense: Prefer ending on scenes that leave a sense of suspense, encouraging viewers to click to find out what happens next.
- Complete Story Arc: It is also acceptable to end where a plot arc is fully wrapped up, giving a satisfying sense of closure while showcasing the highlight.
Output
When generating the output, you should reflect on these requirements and devise an appropriate editing strategy. After careful consideration, select the scenes you believe are suitable for the final cut and include them in a list. Remember to use <result> as the delimiter and present the edited script in the following format:
<result>
[
{"episode":1, "scene_id": 0, "thought": Your justification for choosing this scene. }
{"episode":1, "scene_id": 2, "thought": Your justification for choosing this scene. }
...
]
</result>
Explanation:
- episode: episode number.
- scene_id: scene number within that episode.
- thought: your reasoning and decision for each scene.

{Scene narration of episode 1 $\sim$ T}

Table 10: Instruction template for End2End Editing.

Instruction for End2End Editing with ASR

Role
You are a professional short drama editor with a deep understanding of plot development and audience engagement.
Input
- Drama Title: The name of the short drama.
- Target Audience Gender: The primary gender demographic of the audience.
- Dialogue Information: Dialogue information includes the start time, end time, and the content of each utterance.
Requirements
1. You need to fully understand the content of the input short drama, analyzing the plot and characters to grasp the overall storyline.
2. Based on the scene descriptions, you should edit the drama by preserving the key plot points while ensuring the final cut remains coherent and smooth.
3. Pay special attention to retaining highlight moments within the scenes, as these can significantly enhance the appeal of the edited video. However, always prioritize the overall narrative flow when selecting which highlight moments to include.
I will provide highlight definitions and scoring rules for you.

{Highlight definition and scoring rules (dependent on audience gender)}

Besides highlights, good opening and ending scenes are also crucial.
For Selecting Opening:
You should analyze the plot and consider:
- Audience Engagement: The opening shots should immediately grab attention and draw viewers in.
- Clarity: Avoid clips that rely heavily on prior plot context, such as ones mid-event, which might confuse new viewers.
- Introduction Shots: Shots that introduce characters, setting, or plot premise can be good starting points.
For Selecting Ending:
You should analyze the plot and consider:
- Relevance: Avoid endings that are unrelated to the highlight, as they may transition to a new, less exciting storyline.
- Neutral Endings: If a shot doesn't strongly affect the viewing experience, either way, it can still be chosen as an ending.
- Suspense: Prefer endings that leave a sense of suspense, encouraging viewers to click to find out what happens next.
- Complete Story Arc: It is also acceptable to end where a plot arc is fully wrapped up, giving a satisfying sense of closure while showcasing the highlight.
Output
In the output phase, you can consider these requirements and devise an editing strategy first. Finally, save the segments you deem suitable for editing into a list, and output your decision as a JSON list wrapped in <result> tags, like this:
<result>
[
{"episode":1, "start_time": 0, "end_time": 14, "thought": Your justification for choosing this clip. }
{"episode":1, "start_time": 18, "end_time": 125, "thought": Your justification for choosing this clip. }
...
]
</result>
Explanation:
- episode: episode number.
- start_time: starting time of selected clip.
- end_time: ending time of selected clip.
- thought: your justification for choosing this clip.

{ASR information of episode 1 $\sim$ T}

Table 11: Instruction template for End2End Editing with ASR.