# Supplementary for "Tree-of-AdEditor: Heuristic Tree Reasoning for Automated Video Advertisement Editing with Large Language Model"

**Yuqi Zhang**[1,2], **Bin Guo**[1*], **Nuo Li**[3], **Ying Zhang**[1*], **Shijie Wang**[2], **Zhiwen Yu**[1], **Qing Li**[2*]
[1]School of Computer Science, Northwestern Polytechnical University
[2] Department of Computing, The Hong Kong Polytechnic University
[3] Computation and Artificial Intelligence Innovative College, Fudan University
yuqizhang@mail.nwpu.edu.cn, {guob, izhangying, zhiwenyu}@nwpu.edu.cn
shijie.wang@connect.polyu.hk, qing-prof.li@polyu.edu.hk, linuo@fudan.edu.cn

## 1 Dataset and preprocessing

To further validate the applicability and robustness of ToAE, we introduce ProductAVE, a new dataset containing 152 high-quality videos across diverse scenarios, including home and lifestyle items, toys, food, and beverages. Curating ProductAVE required substantial manual effort to select advertisements with professional cinematographic techniques, such as varied shot compositions, in contrast to low-quality ads with blurred visuals or static product demonstrations. This selection required domain expertise to manually identify suitable videos, making the curation process time-consuming and labor-intensive. During the data preprocessing, 1) We first leverage Transnet [Souvcek and Lokovc, 2020] to split these videos into individual shots. 2) To enhance the robustness for the noisy video input, we design a preprocessing module that filters poor-quality shots using computer vision techniques. This module filters out clips shorter than 1 second, uses Laplacian sharpness to identify blurry clips, and assesses overexposure based on brightness, contrast, and hue variations across different color channels (grayscale, LAB, and HSV). 3) To generate the editing goal for the model input, we exploit the multi-modality model, Videollama [Zhang *et al.*, 2023], to extract the selling points of the product based on its photo by applying the prompt: *Describe the key selling points and features of the product shown in the photo. Include design, materials, functionality, unique aspects, and potential appeal. Ensure the description is clear and highlights the product's market advantages.*

## 2 Additional experiment results

To evaluate ToAE's performance across different LLM capabilities, we conduct experiments using Mistral and GPT-4O-mini, in addition to LLaMA, on both datasets. As Table 5 shows, ToAE outperforms other baselines across varying LLM foundations. Particularly, On GPT-4O-mini, ToAE improves *Stylistic Alignment* by 7.83% on the ProductAVE dataset and 6.79% on the MovingFashion-AVE dataset, compared to IO, CoT, and ToT. In mistral foundation, the *Coherence* is raised by 11.12% with ToAE in ProductAVE. ToAE with visual pruning significantly cuts down the 10% *Jump cut* and *Tonal mismatch*, 20% *Opposite motion* and *Intensity difference* on MovingFashion-AVE. And it also lowers 20% *Opposite motion*, 10% *Intensity difference* and *Tonal mismatch* on ProductAVE. This table also demonstrates the varying abilities of differernt LLMs. The advanced foundation like GPT-4O-mini performs better in high-level dimensions such as *Attractiveness* and *Stylistic Alignment*. While smaller models like Mistral possibly struggle with these higher-level concepts, they still deliver satisfactory performance for real-world applications, especially in resource-constrained environments requiring efficient inference.

## 3 Examples

**Example of Next Shot Selector** Figure 5 presents the specific example of NSS, illustrating the next shot generation and instructions.

**Example of Global Sequence Evaluator** Figure 6 shows the instruction of the Global Sequence Evaluator and specific examples of two evaluated editing plans.

### 3.1 Case study

**Selling Point Discovery and Narrative Organization.** Figure 7 demonstrates the superiority of ToAE in considering selling points and narrative organization, such as first presenting product features then showcasing outfit styling in real scenarios."

**Style Analysis and Visual Coherence.** Figure 8 illustrates that ToAE can analyze the evolving patterns of cinematographic elements in the similar plans and identifies the optimal one that best aligns with customer needs.

## References

[Souvcek and Lokovc, 2020] Tomavs Souvcek and Jakub Lokovc. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

*Corresponding authors

Table 5: Comparison of Different Foundations on MovingFashion-AVE and ProductAVE.

| Dataset | Foundation | Approach | Global Evaluator Score ↑ | | | | Visual Coherence Error ↓ | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Conveyance | Coherence | Attractiveness | Stylistic | Jump cut | Opposite | Intensity | Tonal |
| **MovingFashion-AVE** | Mistral | IO | 85.28 | 61.14 | 76.75 | 59.60 | 0.18 | 0.51 | 0.41 | 0.31 |
| | | CoT | 85.15 | 61.42 | 77.05 | 60.54 | 0.18 | 0.51 | 0.41 | 0.28 |
| | | ToT | 83.98 | 58.72 | 75.90 | 58.42 | 0.18 | 0.54 | 0.47 | 0.34 |
| | | ToAE w/o Visual | 83.99 | 60.29 | 76.28 | 59.21 | 0.15 | 0.52 | 0.43 | 0.33 |
| | | **ToAE** | **88.86** | **67.87** | **80.72** | **66.00** | **0.08** | **0.29** | **0.23** | **0.19** |
| | GPT-4O-mini | IO | 86.42 | 62.29 | 78.59 | 68.52 | 0.20 | 0.55 | 0.46 | 0.33 |
| | | CoT | 80.63 | 62.62 | 81.75 | 72.53 | 0.20 | 0.55 | 0.45 | 0.34 |
| | | ToT | 80.39 | 60.96 | 82.24 | 73.16 | 0.19 | 0.52 | 0.45 | 0.33 |
| | | ToAE w/o Visual | 80.44 | 61.40 | 82.79 | 74.41 | 0.19 | 0.55 | 0.47 | 0.32 |
| | | **ToAE** | 84.62 | **65.50** | **85.27** | **79.95** | **0.11** | **0.37** | **0.29** | **0.21** |
| **ProductAVE** | Mistral | IO | 78.98 | 54.34 | 71.95 | 56.41 | 0.03 | 0.57 | 0.24 | 0.55 |
| | | CoT | 79.11 | 55.46 | 73.42 | 56.45 | 0.03 | 0.55 | 0.25 | 0.54 |
| | | ToT | 77.76 | 53.99 | 74.61 | 56.49 | 0.03 | 0.55 | 0.19 | 0.59 |
| | | ToAE w/o Visual | 78.52 | 54.12 | 74.41 | 56.74 | 0.03 | 0.59 | 0.21 | 0.59 |
| | | **ToAE** | **84.89** | **65.11** | **79.97** | **65.05** | 0.02 | **0.43** | **0.11** | **0.45** |
| | GPT-4O-mini | IO | 79.47 | 61.45 | 82.89 | 73.95 | 0.02 | 0.58 | 0.23 | 0.56 |
| | | CoT | 79.01 | 63.03 | 83.29 | 73.03 | 0.02 | 0.57 | 0.21 | 0.54 |
| | | ToT | 79.78 | 60.29 | 81.32 | 71.47 | 0.03 | 0.58 | 0.18 | 0.53 |
| | | ToAE w/o Visual | 80.07 | 60.00 | 82.38 | 73.77 | 0.02 | 0.59 | 0.19 | 0.56 |
| | | **ToAE** | **83.52** | **64.23** | **87.18** | **79.30** | 0.02 | **0.43** | **0.12** | **0.44** |

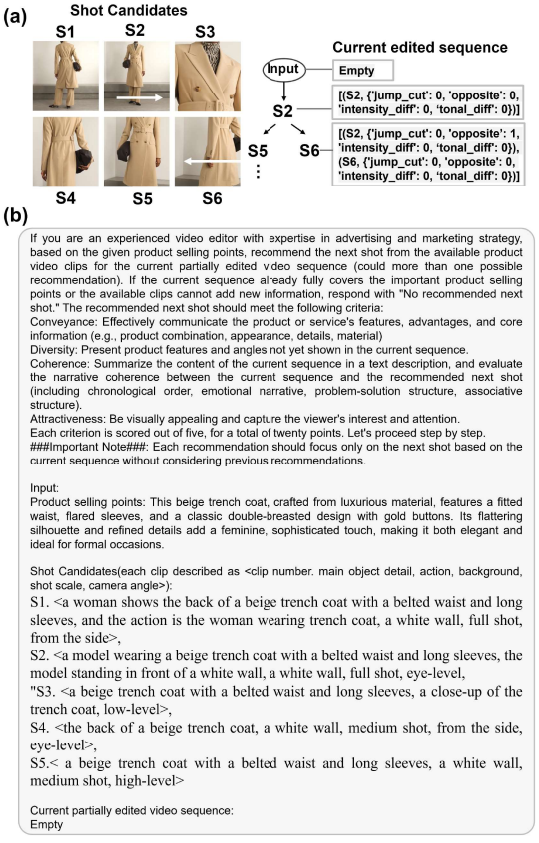


Figure 5: The specific example of Next shot Selector. (a) The process of next shot generation. S2 is chosen as the start shot to establish the whole picture of the coat. The possible next shot, S5, provides a diverse scale, focusing on the front of the coat and highlighting selling points like the waist and buttons. Another possible option, S6, shows the side of the coat to emphasize the sleeves while forming an opposite motion with S2. (b) The instruction for the Next Shot Selector at the initial state.

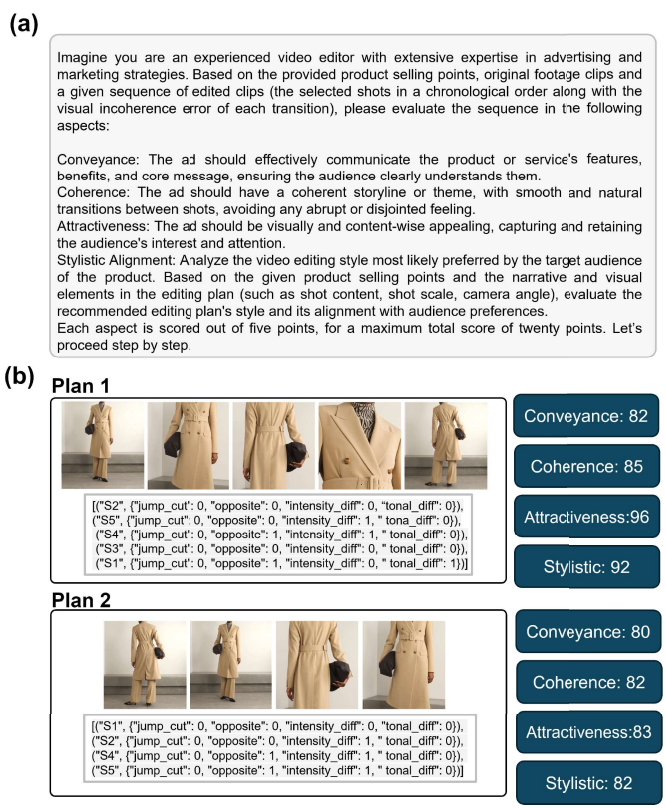


Figure 6: The specific example of Global Sequence Evaluator. (a) the instruction of the Global Sequence Evaluator for each editing plan, (b) Evaluation of two editing plans. The first plan obtains higher scores in *Attractiveness* and *Stylistic Alginment* due to the reason given by GSE, namely, "the use of medium shots, close-ups, and side angles creates a sophisticated and luxurious feel, fitting for a high-end fashion brand."

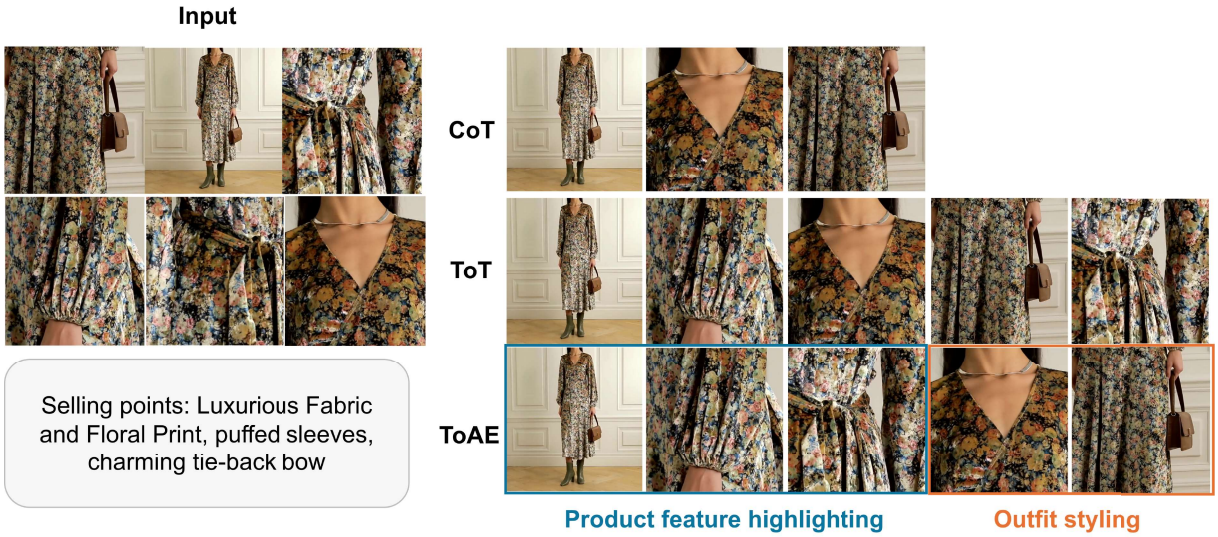


Figure 7: Case study of comparative results on selling points discovery and narrative organization. ToAE highlights more selling points than CoT, covering all essential features. It also employs an effective narrative technique: first showcasing the product features (fabric, sleeves, and tie-back bow), then presenting outfit styling suggestions. This clear and logical progression enhances the persuasive impact of the ad, making the product's advantages more memorable and relatable to the audience.


Figure 8: Case study of comparative results on style analysis and visual coherence. Although these two plans both cover the selling points for the parka, the different evolving patterns of shot scale present distinct editing styles. ToT starts with a full shot, then shifts to a close-up to focus on product details, returns to a full shot, and progresses with medium, close-up, and medium shots. It is more dynamic but can feel disjointed due to frequent shifts between wide and close-ups, potentially disrupting narrative continuity. ToAE offers a more balanced and natural progression. It begins with a full shot to establish the overall environment, followed by a close-up to highlight product details, then medium shots to expand on the narrative, returning to a close-up for emphasis, and ending with a full shot. This organization way effectively guides viewers from the overall scene to the details and back, creating a coherent narrative and a smoother visual experience.