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

Abstract

Video advertising has become a popular marketing strategy on e-commerce platforms, requiring high-level semantic reasoning like selling point discovery, narrative organization. Previous rule-based methods struggle with these complex tasks, and learning-based approaches demand large datasets and high training costs. Recently, Large Language Models have opened incredible opportunities for advancing intelligent video advertisement editing. However, IO and CoT struggle to adapt to the non-linear thinking hierarchy of video editing, where editors iteratively select shots or revert them to explore potential editing solutions. While ToT (Tree-of-Thought) offers a conceptual structure that mirrors this hierarchy, it falls short in aligning with effective video advertising strategies and lacks robust fact-checking mechanisms. To address these, we propose a novel framework, Tree-of-AdEditor (ToAE), which constructs a reasoning tree to mimic human editors, and incorporates domain-specific theories and heuristic fact-checking to identify optimal editing solutions. Specifically, motivated by effective advertisement principles, we develop a "local-global" mechanism to guide LLM in both the shot level and sequence level decision-making. We introduce a visual incoherence pruning module to provide external heuristic fact-checking, ensuring visual attractiveness and reducing computation costs. Quantitative experiments and expert evaluation demonstrate the superiority of our method compared to baselines.

1 Introduction

Video advertising has become an increasingly popular marketing strategy to capture consumers' attention and drive purchase behavior on e-commerce platforms like Taobao and Amazon [Liu and Yu, 2023]. Crafting effective video advertisements requires adherence to several fundamental principles, such as highlighting abundant product selling points, maintaining a coherent narrative, and aligning with the preferred editing styles of the target consumers [Armstrong, 2010]. Creating such video advertisements necessitates both editing expertise and marketing acumen, and it is time-consuming to meticulously select and sequence shots from large footage repositories. All these make large-scale, effective product promotion on e-commerce platforms unfeasible.

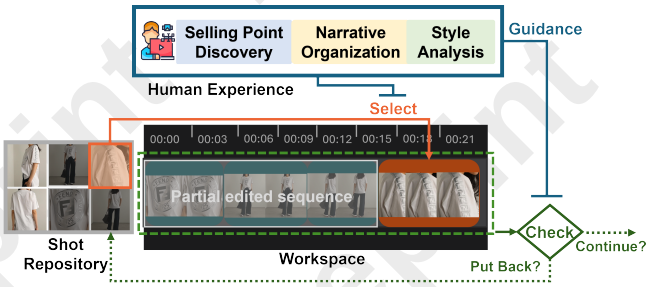

Figure 1: Manual editing process. Editors iteratively select the next shot from the shot repository or put it back, considering the selling point, narrative, and style. Repeat until the editing goal is achieved.

Previous works on video advertisement editing mostly rely on predefined rules to constrain the visual quality and coherence [Lin *et al.*, 2021; Liu *et al.*, 2021; Liu *et al.*, 2019]. While these fixed patterns enhance the overall visual attractiveness, this closed-world setting struggles with narrative organization and selling points discovery. Other learning-based methods try to mine professional knowledge from well-edited videos, like cut-trigger patterns [Pardo *et al.*, 2021] and long-range associations among shots [Argaw *et al.*, 2023]. However, this branch of studies lacks an explicit explanation and requires massive datasets and training costs.

Recently, Large Language Models (LLMs) have exhibited impressive zero/few-shot reasoning capabilities across various planning tasks [Huang *et al.*, 2024]. The high-level semantic understanding and abundant domain knowledge of LLMs provide a golden opportunity to advance intelligent video advertisement editing. Typically, video advertisement editing involves selling point discovery, narrative organization, and style understanding. It is challenging for vanilla input-output (IO) prompting to consider all components in only one step. Chain-of-Thought (CoT) [Wei *et al.*, 2022;

*Corresponding authors
Supplementary: github.com/GeniusEditor/Tree-of-AdEditor

Kojima *et al.*, 2022; Wang *et al.*, 2022] can decompose hard problems into several intermediate steps and infer the final answer step by step along this linear thinking chain. Although the CoT achieves significant improvement, it hardly adapts to *non-linear and iterative thinking hierarchy* of the video editing process, *where editors iteratively try different shots or put them back to explore multiple potential editing paths, ultimately fulfilling the creation purpose*. For the improved structure based on CoT, Tree-of-Thoughts (ToT) [Yao *et al.*, 2024] builds a reasoning tree to globally explore and self-evaluate each branch of thought, which closely mirrors the manual editing process illustrated above. However, ToT still faces two significant challenges in video advertisement editing: **1) Alignment with Effective Video Advertising Principles:** To align the LLMs with the goal of effective video advertisement (e.g., informative, coherent, and stylistic) [Armstrong, 2010; Scott Armstrong, 2011; Valentini *et al.*, 2018; Liu and Yu, 2023; Bowen, 2017], how to design domain-specific prompts to guide the node generation and optimal path selection in the tree inference structure is a challenging issue. **2) External Fact-Checking for Sub-Step:** ToT simply conducts self-evaluation without any external knowledge, resulting in unfeasible and cost-ineffective searches.

To address the above challenges, we propose a novel framework, Tree-of-AdEditor (ToAE), which constructs a reasoning tree that can iteratively look ahead ("select") and backtrack ("put back") to precisely imitate human editors. Given the video footage of the product and the brief description of the editing goal, ToAE achieves systematic planning and identifies the optimal editing solution regarding informativeness, coherence, and stylistic preference alignment with target consumers (also easily adaptable to other criteria if needed). Specifically, to extract the essential information for the editing task and address the modality misalignment between multi-shot data and LLM-understandable format, we first leverage an explicit shot semantic representation module. This module encodes each shot, including content details and cinematographic elements like shot scale. Then, we develop a 'local-global' mechanism that precisely guides the LLMs through both shot-level decisions and overall sequence evaluation. The Local Next Shot Selector recommends a good next shot as a tree node, while the Global Sequence Evaluator holistically assesses each potential editing plans derived from the reasoning tree. Finally, to ensure visual attractiveness, we leverage the heuristics of visual coherence errors to provide external fact-checking, which effectively cuts down the the computation cost on the visually abrupt nodes. To conclude, our contributions are summarized below:

- We propose Tree-of-AdEditor (ToAE), a novel framework that uses a reasoning tree to mimic manual video editing with LLM knowledge and visual heuristics, to our knowledge, the first of its kind.
- Motivated by Effective Video Advertising Principles, we design a "local-global" mechanism that guides the LLM at both the shot level and sequence level decision-making.
- Quantitative experiments on MovingFashion and our novel dataset, ProductAVE, demonstrate the effectiveness and efficiency of ToAE compared to advanced baselines. Human evaluations further validate ToAE's superiority, with 43% of professionals favoring it over 32% for ToT.

2 Related Work

2.1 Video Advertisement Editing

The video advertisement editing process consists of shot selection and sequencing, both essential for creating an informative and engaging visual narrative. Many methods rely on manually annotated datasets to train classifiers for shot selection, such as narrative importance [Liu *et al.*, 2019]. However, these approaches incur high costs and rely on subjective and hard-to-standardize annotations. For shot sequencing, existing methods often relies on predefined rules [Lin *et al.*, 2021; Liu *et al.*, 2021; Tang *et al.*, 2022; Liu *et al.*, 2019; Galvane *et al.*, 2015; Leake *et al.*, 2017; Arev *et al.*, 2014; Wang *et al.*, 2019], and they are hardly imitate complicated open-domain editing techniques and struggle with high-level tasks like narrative organization, selling point discovery, and style understanding. Some learning-based methods in the movie domain [Argaw *et al.*, 2022; Pardo *et al.*, 2021; Argaw *et al.*, 2023] learn professional editing knowledge like cut-trigger patterns [Pardo *et al.*, 2021]. However, they lack interpretability and require large datasets and significant training costs.

2.2 Planning with LLMs

Recently, Large Language Models (LLMs) have demonstrated impressive reasoning abilities in tasks like mathematics, creative writing, and multi-hop question answering [Huang *et al.*, 2024; Hazra *et al.*, 2024; Zhang *et al.*, 2024; Huang *et al.*, 2022]. Various methods, such as few-shot demonstrations and refinement mechanisms [Madaan *et al.*, 2024; Gou *et al.*, 2023; Shinn *et al.*, 2024], and accessing external knowledge [Guan *et al.*, 2023; Press *et al.*, 2022], have been proposed to enhance their capabilities. One distinct branch constructs the thought structure to solve the task little by little. The Chain-of-Thoughts (CoT) method [Wei *et al.*, 2022; Kojima *et al.*, 2022] reasons along a linear thought chain, and CoT-SC [Wang *et al.*, 2022] was introduced to enhance robustness. While CoT-like methods struggle with the non-linear nature of video editing and lack a fact-checking mechanism to prevent error propagation. Tree-of-Thoughts (ToT) [Yao *et al.*, 2024] constructs a reasoning tree that iteratively looks ahead and backtracks, mimicking human editors' cognitive process. However, general ToT lacks reliable fact-checking and effective guidance to induce shot-level decision and overall sequence-level planning for unique video advertisement editing tasks.

3 Preliminary

3.1 Characteristics of Video Advertisement Editing

Video advertisement editing involves selecting and arranging footage to highlight product features, engage the audience, and drive purchases. Editors typically select shots iter-

atively, organize them into a coherent narrative, and add captions, music, or special effects [Bowen, 2017]. Among these steps, shot selection and assembly are crucial, as they shape the narrative and determine the ad's resonance with the target audience. This paper focuses on optimizing the iterative process of shot selection and assembly, with advanced effects and post-production to be explored in future work.

Motivated by the theories in advertisement marketing and video editing [Armstrong, 2010; Scott Armstrong, 2011; Valentini *et al.*, 2018; Liu and Yu, 2023; Bowen, 2017], we summarize three pivotal characteristics of effective ad video advertisement editing:

- **Informativeness.** The advertisement should showcase key product selling points, functional features, and usage scenarios to provide customers with the necessary information to make informed decisions. Irrelevant or non-essential content should be avoided to keep the message concise and easily understood.
- **Coherence.** A coherent narrative ensures a smooth storyline, reducing the cognitive load of viewers and preventing confusion. Smooth shot transitions, such as scale changes and angle shifts, allow each shot to flow naturally into the next, while also enhancing visual variety to sustain viewer interest and anticipation.
- **Attractiveness.** Visual quality, narrative structure, cutting rhythm significantly impact attractiveness. Aligning these elements with target audience preferences further boosts appeal and encourages purchase behaviors.

While our framework emphasizes these aspects, it is highly adaptable to incorporate other relevant factors if needed.

3.2 Problem Formulation

Motivated by these fundamental principles, we aim to generate informative, coherent and stylistic video advertisement A, given the shot repository $S = \{s_1, .., s_n\}$ and editing goal description (e.g., selling points) D provided by product business. Each shot s_i is cut from the original recordings, representing a unit that captures an uninterrupted action, event, or scene. We formulate this task as an **optimal path-finding problem in the tree search**. The "search" and "backtrack" operations in the tree construction can precisely mimic the "select" and "put back" actions of human editors. Specifically, our Tree-of-Editor (ToAE) framework takes the LLM p_θ, shot repository S, editing goal description D, and instructions P as input to construct a reasoning tree $\mathcal{T}$ and to find the optimal editing solution A^*. Here we denote $\mathcal{T} = (\mathcal{N}, \mathcal{E})$ as the reasoning tree. Each tree node $n \in \mathcal{N}$ represents a selected shot from the shot repository, with edges $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ from parent nodes to child nodes indicating shot transitions. The reasoning path A_i within the tree is represented as a series of connected nodes, beginning at the root and extending to a leaf node. The optimal editing solution A^* is derived from the set of all possible editing plans $\{A_i\}$ in the searched tree $\mathcal{T}(S, D, P, p_\theta)$, aligning best with the editing goal D. The ultimate goal is formulated as:

$$A_i = \{n_{\text{root}}, \ldots, n_{\text{leaf}}\}, (n_j \to n_{j+1}) \in \mathcal{E},$$
$$A^* = \arg \max_{A_i \in \mathcal{T}(S,D,P,p_\theta)} \mathcal{U}(A_i; D) \quad (1)$$

where function $\mathcal{U}(A_i; D)$ evaluates video advertisements based on informativeness, coherence, and attractiveness.

4 Methodology

4.1 Overview

As Figure 2 shows, our ToAE framework consists of: 1) ***Explicit Shot Semantic Representation*** encodes each video clip as textual descriptions of its content details and cinematographic metrics. This serves as the node representation in the tree reasoning, bridging video-text modality gaps between the multi-clips data and LLM-understandable formats. 2) ***Local-Global Guidance*** navigates the LLM in shot-level and sequence-level decision-makings. ***Local Next Shot Selector*** integrates selling points and narrative structure to determine next shots as child nodes. And ***Global Sequence Evaluator*** assesses all potential plans derived from the reasoning tree, focusing primarily on stylistic alignment with the target audience and other content-related attributes. 3) To enhance visual attractiveness, ***Visual Incoherence Pruning*** provides external fact-checking to identify visual coherence errors between newly selected shots and the historical sequence. If the transition is visually abrupt or incoherent, the new branch is pruned. Finally, the plan with the highest score from the reasoning tree is selected as the optimal solution. The last subsection details the DFS search process and inference cost.

4.2 Explicit Shot Semantic Representation

LLMs demonstrate superior performance across various NLP tasks. However, emerging multi-modal LLMs like DALL-E [Ramesh *et al.*, 2021] and Videollama [Zhang *et al.*, 2023], struggle with reasoning in the video domain, particularly when handling multiple clips as input. Video ad editing also requires the understanding of content details and cinematographic elements, making the above methods barely applicable. To enable effective LLM inference in multi-clip editing, we employ a video-text model to encode the content and cinematographic elements of each clip. They serve as node representations in the reasoning tree, effectively bridging the gap between video inputs and LLM-compatible formats.

Specifically, editors usually focus on the content of each shot, including the main object, the action, and the background, and the cinematographic elements like shot scale and camera angle [Bowen, 2017; Smith, 2006]. Hence, the textual description of each shot s_i is represented as $s_i^d = R(s_i) =$*<main object detail, action, background, shot scale, camera angle>*, and $R(\cdot)$ is the text-video model. We design explicit prompt to encode the representation as follows: *This is a shot from a product advertisement video of [product name]. What does this shot describe (include main object detail, action, background, shot scale, camera angle)?*

4.3 Local-Global Guidance of Effective Advertising

To align with effective video advertising principles, we design the "local-global" guide to navigate the LLM for node

defined as the proportion of the main object in the whole screen, e.g., the close-up, the full shot.

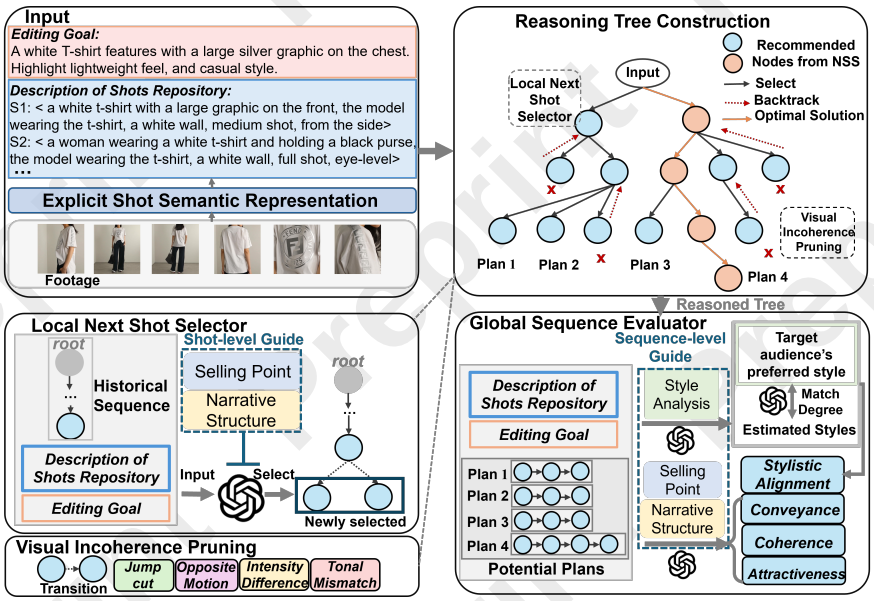

Figure 2: Our ToAE framework for effective video advertisement editing. Incorporate "local-global" guidance (Local Next Shot Selector and Global Sequence Evaluator) and external visual heuristic fact-checking to find optimal editing plan.

generation and path evaluation, namely, *Local Next Shot Selector* and *Global Sequence Evaluator*.

Local Next Shot Selector

In the shot selection and assembly process, editors usually consider two primary aspects: the shot itself and its transition from the historically edited shot sequence. Motivated by the principles of effective advertising, a good video should showcase product features, avoid irrelevant content, attract viewers, and maintain narrative coherence. Hence, we propose *Conveyance, Diversity, Attractiveness* for the LLM to select shots that effectively convey the selling points and present certain differences from other shots in the ad. And *Coherence* evaluates the narrative coherency between the candidate shot and historical sequence. According to the above criteria, the next shot selector $N(\cdot)$ recommends feasible subsequent nodes $n_{t+1}^{(1...k)}$ based on the historical path $(n_1^{root}, n_2^{i2}, ..., n_t^{it})$ from the root node to the current node, which is calculated as:

$$n_{t+1}^{(1...k)} = N(p_\theta, (n_1^{root}, n_2^{i2}, ..., n_t^{it}), S^d, D, P^{select}) \quad (2)$$

Where n_t^{it} denotes the it-th node at the t level, k represents the number of recommended next nodes, and P^{select} is the prompt for the next shot selector.

Global Sequence Evaluator

The Next Shot Selector guides shot selection, but controlling style such as cutting rhythm and narrative structure, remains challenging at the shot level. Therefore, we propose the Global Sequence Evaluator to holistically assess each potential editing plan from the reasoning tree, focusing on stylistic alignment and content-related aspects at the sequence level.

According to video editing theories [Bowen, 2017; Choudhary *et al.*, 2019], style can be formulated as the distribution of shot attributes and their evolving patterns over time. For instance, a transition from a full shot to a close-up can intensify

Figure 3: Visual coherence error

emotions. Hence the style of an editing plan A_i is computed as: $L(A_i) = p_\theta(\{n_{\text{root}}, \ldots, n_{\text{leaf}}\}), n_j$ =<*main object detail, action, background, shot scale, camera angle*>, $A_i \in \mathcal{T}$. To evaluate *Stylistic Alignment*, we first induce LLM to infer the target audience's preferred style $\hat{L} = p_\theta(S^d, D)$ and then let LLM evaluate the style of the given editing plan $L(A_i)$, finally infer its match degree with target users' preferred style $p_\theta(\hat{L}, L(A_i))$. In addition, inspired by the principles of effective advertisement practice, we also summarize *Conveyance, Coherence, Attractiveness* altogether. The process of the global sequence evaluator is computed as follows:

$$\{Score_i^z\} = G(p_\theta, A_i, S^d, D, P^{eval}) \quad (3)$$

where $z \in \{$*Conveyance, Coherence, Attractiveness, Stylistic Alignment*$\}$.

4.4 Visual Incoherence Pruning

The Local-Global Guidance relies solely on textual descriptions, often causing visually abrupt transitions that impair

narrative clarity and aesthetic appeal. To address this, we introduce the Visual Incoherence Pruning module, which fact-checks visual coherence at each sub-step in the reasoning tree, focusing on transitions between newly selected shots and the historical sequence. This module ensures visual cohesion, reduces unnecessary exploration, and lowers inference steps.

Here we focus on four common types of errors for video advertisement scenarios, as shown in Figure 3:

- *jump cut:* appears when two subsequent shots present very similar content from slightly different angles or positions, leading to a noticeable gap. $C_{jc}(\cdot)$ is calculated by the similarity of SURF keypoints [Bay *et al.*, 2008].
- *opposite motion:* $C_{opp}(\cdot)$ is calculated by the angle difference of these motion vectors. The motion feature is derived from the optical flow tracking [Sun *et al.*, 2010].
- *intensity difference:* $C_{int}(\cdot)$ computes as the speed difference of motion features.
- *tonal discontinuity:* $C_{tonal}(\cdot)$ compares the similarity of the HSV color histograms.

We use these heuristics to fact-check the visual transition from the last shot of the historical sequence and newly selected shots (from *Next Shot Selector*). If the visual incoherence error exceeds the threshold ($C_{prune}(n_{t+1}^{(q)}) > \varepsilon$), this branch will be pruned.

$$C_{prune}(n_{t+1}^{(q)}) = \sum_{l} f_l(C_l(I(n_t^{it}), I(n_{t+1}^{(q)}))),$$
$$l = \{jc, opp, int, tonal\}, q = 1, .., k \tag{4}$$

where $f_l(\cdot)$ judges whether the corresponding errors exists. 0 for "not exist", 1 for "exist", and function $I(\cdot)$ maps a shot's textual description to the corresponding video shot.

4.5 Cost-Efficient Optimal Path Search Algorithm

Algorithm 1 outlines the depth-first search (DFS) process for constructing the ToAE reasoning tree, mimicking manual video editing by deeply exploring potential shots and backtracking. Starting with an empty sequence seq_0, next shot selector $N(\cdot)$ generates candidate nodes and explore them from high to low confidence until a stop condition is met (either the tree size exceeds $num > M$ or no candidates remain). For each newly selected node n', visual incoherence with the historical sequence is evaluated. If the coherence error exceeds threshold ε, the subtree is pruned and backtracks. Otherwise, this new node n' will be added to the current sequence seq and exploration continues. Finally, all potential plans $\{A_i\}$ from the reasoned tree are evaluated by global sequence evaluator in terms of the above four criteria, with the highest-scored plan chosen as the optimal solution A^*.

$$A^* = \arg \max_{A_i \in \mathcal{T}(S,D,P,p_\theta)} \mathcal{U}(A_i; D) = \arg\max_i \sum_z Score_i^z \tag{5}$$

where $P = (P^{select}, P^{eval}, C_{prune})$.

Inference Cost. *The computation complexity of ToAE is $O((p * n)^d)$, where n is the number of shot candidates and d denotes the depth of the reasoning tree.*

Proof: Each node is selected without replacement (i.e., once a shot is selected, it is no longer available unless backtracked). Initially, there are n possible choices at the first level, $n-1$ at the second level, and so on, resulting in a tree size of $O(n(n-1)(n-2)(n-d))$. When visual pruning is applied and only a fraction p^d of the branches is retained at each level d, thereby the total nodes of reasoning tree can be expressed as $\prod_{d=0}^{D} p^d(n-d)$. Thus, the total complexity of ToAE is $O((p*n)^d)$. □

Compared to the complexity for naïve permutation $O(n!)$, ToAE cuts down much computation costs and becomes computationally feasible for larger values.

Algorithm 1 Tree of Editor-DFS

procedure ToAE(p_θ, S, D, P, M, ε)
  $S^d \leftarrow R(S, P^{encode})$
  $seq_0 \leftarrow empty$
  $\{A_i\} \leftarrow DFS(seq_0, S^d, D, 0)$
  $A^* \leftarrow \arg\max_i \sum_z G(p_\theta, A_i, S^d, D, P^{eval})$
  return A
end procedure
procedure DFS(seq, S^d, num)
  $nexts \leftarrow N(p_\theta, seq, S^d, D, P^{select})$
  if $num > M$ or $nexts$ is empty **then**
    $A \leftarrow record(seq)$
    return A
  end if
  for $n' \in nexts$ **do**
    sorted candidates
    $q_1 \leftarrow I(seq[last])$, $q_2 \leftarrow I(n')$
    if $C_{prune}(q_2) > \varepsilon$ **then**
      prune
    else
      update $seq' \leftarrow assembly(seq, n')$
      $num \leftarrow num + 1$
      DFS(seq', S^d, num)
    end if
  end for
end procedure

5 Experiment

5.1 Dataset and Preprocessing

MovingFashion-AVE: MovingFashion [Godi *et al.*, 2022] includes fashion video advertisements from the e-commerce platform Net-A-Porter . We create 458 editing tasks ranging from 3 to 6 shots in the fashion advertisement scenario.
ProductAVE (Ad Video Editing): comprises 152 Taobao video advertisements across various categories, including furniture, kitchenware, home decor, toys, etc. These videos are typically longer and more challenging than fashion ads, with lengths ranging from 3 to 10 shots after preprocessing. More details are explained in Supplementary.

5.2 Metrics

We employ the LLMs scorings (*Conveyance, Coherence, Attractiveness, Stylistic Alignment*), and Visual heuristics (*Jump cut, Opposite Motion, Intensity Difference, Tonal Mismatch*)

https://www.net-a-porter.com
https://www.taobao.com/

Dataset	Approach	Global Evaluator Score ↑				Visual Coherence Error ↓			
		Conveyance	Coherence	Attractiveness	Stylistic	Jump cut	Opposite	Intensity	Tonal
MovingFashion-AVE	IIC	89.21	58.53	79.35	81.86	0.20	0.56	0.44	0.30
	LMP	89.03	58.56	79.37	82.13	0.19	0.56	0.47	0.33
	IO	89.59	64.41	82.18	86.07	0.20	0.56	0.45	0.32
	CoT	89.67	63.97	81.48	86.03	0.19	0.53	0.47	0.35
	ToT	89.78	65.38	83.06	87.00	0.19	0.55	0.45	0.34
	ToAE w/o Visual	89.91	65.82	83.41	86.13	0.19	0.54	0.46	0.33
	ToAE	89.81	**70.53**	**86.10**	**89.63**	**0.11**	**0.37**	**0.28**	**0.22**
ProductAVE	IO	86.58	55.26	76.58	78.29	0.20	0.60	0.20	0.60
	CoT	87.24	55.53	76.45	79.47	0.03	0.62	0.21	0.52
	ToT	86.67	56.11	77.22	78.89	0.03	0.61	0.22	0.56
	ToAE w/o Visual	87.76	57.76	78.57	79.80	0.03	0.60	0.21	0.56
	ToAE	**88.61**	**59.44**	**79.44**	**81.67**	0.03	**0.45**	**0.11**	**0.42**

Table 1: Quantitative results comparing with baselines

Approach	Global Evaluator Score ↑				Visual Coherence Error ↓			
	Conveyance	Coherence	Attractiveness	Stylistic	Jump cut	Opposite	Intensity	Tonal
ToAE w/o encoding	89.01	64.36	81.92	87.06	0.12	0.36	0.28	0.22
ToAE w/o GSE	89.81	65.89	83.23	87.17	0.16	0.51	0.39	0.31
ToAE w/o Visual	89.91	65.82	83.41	86.13	0.19	0.54	0.46	0.33
ToAE	89.81	**70.53**	**86.10**	**89.63**	**0.11**	**0.37**	**0.28**	**0.22**

Table 2: Ablation results

for evaluation. Higher scores in the first four dimensions and lower visual incoherence errors indicate better outcomes.

5.3 Implementation Details

We harness the llama3.1:70b [Touvron *et al.*, 2023] deployed in the local device for our experiment (to verify the applicability of ToAE in different LLM foundation, we also compare the performance in Mistral and GPT-4o-mini in the Supplementary). We employ a multi-modal large language model, VideoLLama [Zhang *et al.*, 2023], to encode the raw video footage into the textual description. Our initial observations suggest that VideoLLama extracts more detailed and nuanced information, making it better suited to our application scenario compared to VideoChat [Li *et al.*, 2023] and VideoChatGPT [Maaz *et al.*, 2023]. For the next shot selector, the number of generated answers is determined by LLMs. We limit the DFS process to a maximum of 100 inference steps to balance the performance and efficiency.

5.4 Baselines

Inter-Intra Contrastive (IIC) [Tao *et al.*, 2020] and Long-range Multimodal Pretraining (LMP) [Argaw *et al.*, 2023] both utilize their learned representation for shot sequencing (retrained on both two datasets using 3-fold cross-validation). IO generates directly with LLM. CoT (Chain-of-Thought) [Kojima *et al.*, 2022] generates in a linear chain based on solely textual information. ToT (Tree-of-Thought) [Yao *et al.*, 2024] infers in a reasoning tree based on only textual information. ToAE w/o Visual generates without visual checking.

5.5 Comparison with Baselines

Global Evaluator Scoring. Table 1 shows that ToAE achieves the best performance across both datasets, showcasing its effectiveness and adaptability in diverse scenarios. Compared to learning-based methods like IIC and LMP, this highlights the limitations of small models and the advantage of using domain knowledge from LLMs for automatic video editing. ToT outperforms IO and CoT, demonstrating the effectiveness of tree-structured reasoning. Comparing ToAE w/o Visual with ToT underscores the importance of "local-global" guidance in tree construction. Table 3 deomstrates its consistent performance on different LLM foundations, highlighting its robustness with varying model capabilities. Additional comparisons can be found in the Supplementary.

Approach	Conveyance	Coherence	Attractiveness	Stylistic
Mistral				
IO	85.28	61.14	76.75	59.60
CoT	85.15	61.42	77.05	60.54
ToT	83.98	58.72	75.90	58.42
ToAE w/o Visual	83.99	60.29	76.28	59.21
ToAE	**88.86**	**67.87**	**80.72**	**66.00**
GPT-4O-mini				
IO	86.42	62.29	78.59	68.52
CoT	80.63	62.62	81.75	72.53
ToT	80.39	60.96	82.24	73.16
ToAE w/o Visual	80.44	61.40	82.79	74.41
ToAE	84.62	**65.50**	**85.27**	**79.95**

Table 3: Results on MovingFashion-AVE with different foundations

Visual Coherence Error. Given that error values vary with different footage, we focus on the relative reduction. Table 1 demonstrates that ToAE achieves significant improvements, reducing *Jump Cut* error, *Tonal Mismatch*, and *Intensity Difference* by at least 10 % and minimizing *Opposite Camera Motion* by at least 15 % compared to IO, CoT, and ToT.

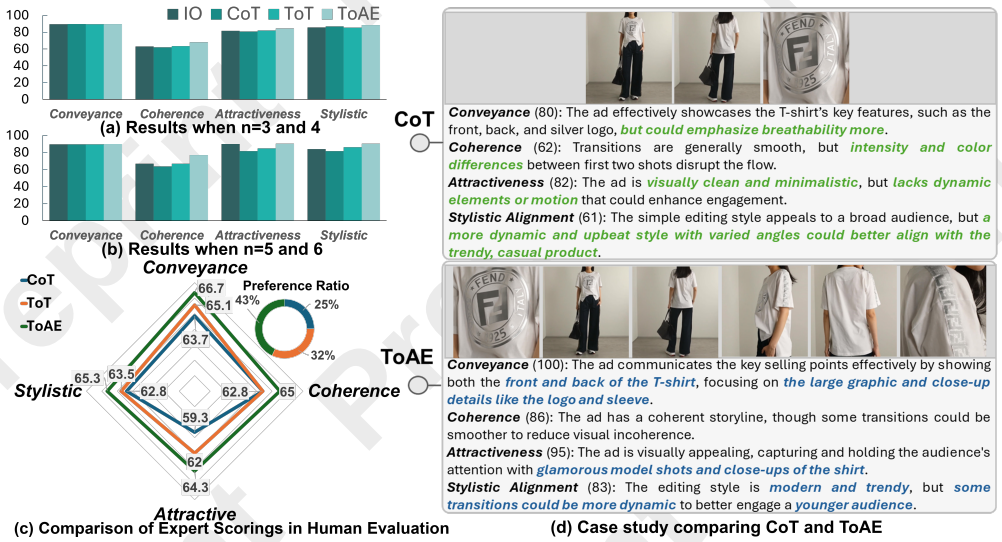

Figure 4: **(a)(b) Comparision in different inputing size. (c) Human Evaluation results.** ToAE achieves the highest scores across all dimensions and is preferred by 43% of professionals, surpassing CoT (25%) and ToT (32%). **(d) Example of generated Ad.** ToAE emphasizes more key product features and considers dynamic scale transitions, aligning with the young audience's preference for modern trends.

Effects of the Length of Candidate Shots. As the shot repository grows, the algorithm needs to consider more information and combinations, making inference challenging. Figure 4 (a)(b) shows that ToAE achieves higher performance even under more challenging settings when $n = 5, 6$. ToAE achieves at least 10.2%, 6.4%, and 5.2% improvement regarding *Coherence*, *Attractiveness*, and *Stylistic* compared to IO and CoT. The improvement may stem from a richer pool, enabling more diverse shot assembly. This demonstrates the effectiveness and robustness of ToAE in typical-length e-commerce ads, highlighting its real-world applicability.

Approach	n=3	n=4	n=5	n=6
Naive Enumeration	15	64	325	1956
ToAE w/o Visual	4.39	9.17	19.05	41.57
ToAE	**4.28**	**7.90**	**15.25**	**30.89**

Table 4: Inference steps in varying input size.

Efficiency of ToAE. Table 4 shows the exponential growth in computational cost for naive enumeration, with inference steps reaching 64 for an input size of only 4. ToAE w/o Visual demonstrates the effectiveness of LLM domain knowledge and 'local-global' guidance in avoiding unlikely exploration. Additionally, ToAE outperforms ToAE w/o Visual by reducing approximately 10 steps when $n = 6$. It validates visual fact-checking's effectiveness in reducing abrupt branches and lowering computation costs. Despite the effectiveness of ToAE in short video ad production, it still faces challenges with longer ads, which will be addressed in the future work.

5.6 Ablation Study

As Table 2 shows, ToAE w/o encoding relies on general video-text LLM descriptions instead of leveraging explicit cinematographic details, which are critical for ensuring coherence, attractiveness, and stylistic alignment. ToAE w/o GSE (Global Sequence Evaluator) generates less coherent and stylistically aligned video advertisements. The comparison of ToAE w/o Visual with ToAE highlights the importance of visual heuristic fact-checking in avoiding coherence errors.

5.7 Human Evaluation

To validate our method in real-world scenarios, we conducted a human evaluation with six ***highly qualified professionals***, each with extensive experience in video editing and advertising, including some with ***over 15 years of expertise***. Specifically, participants are asked to evaluate generated video advertisements for ten products, each with three generated ads by CoT, ToT, and ToAE. The evaluation scale ranges from 1 to 10, with 1 indicating poor and 10 indicating excellent. Additionally, participants are required to select their preferred advertisement for each product. The results of human evaluation are shown in Figure 4 (c). ToAE outperforms other methods, achieving the highest scores across all dimensions and being favored by 43% of professionals, compared to 25% for CoT and 32% for ToT. One example is shown in Figure 4 (c), and more examples could be found in Supplementary.

6 Conclusion

This paper presents the first exploration of leveraging LLMs for intelligent video editing, aiming to generate engaging short ecommerce advertisements based on user-provided video footage and editing goals. We propose a novel Tree-of-AdEditor (ToAE) framework that constructs a reasoning tree to mimic the cognitive process of human editors. This frame work integrates fundamental principles in video editing and advertising marketing and external visual heuristics to address the significant challenges of adapting ToT in the video advertisement editing. Future research could explore integrating music and special effects or addressing the challenges of producing long advertisements.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No. 62025205), the Hong Kong Research Grants Council under General Research Fund (No. 15200023), Research Impact Fund (No. R1015-23), and National Natural Science Foundation of China (No. 62432007, No. 62272390).

References

[Arev *et al.*, 2014] Ido Arev, Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir. Automatic editing of footage from multiple social cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014.

[Argaw *et al.*, 2022] Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing. In *European Conference on Computer Vision*, pages 201–218. Springer, 2022.

[Argaw *et al.*, 2023] Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, and Fabian Caba Heilbron. Long-range multimodal pretraining for movie understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13392–13403, 2023.

[Armstrong, 2010] J Armstrong. *Persuasive advertising: Evidence-based principles*. Springer, 2010.

[Bay *et al.*, 2008] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.

[Bowen, 2017] Christopher Bowen. *Grammar of the Edit*. Routledge, 2017.

[Choudhary *et al.*, 2019] Priyankar Choudhary, Neeraj Goel, and Mukesh Saini. A multimedia based movie style model. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 72–77. IEEE, 2019.

[Galvane *et al.*, 2015] Quentin Galvane, Rémi Ronfard, Christophe Lino, and Marc Christie. Continuity editing for 3d animation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 753–761. AAAI Press, 2015.

[Godi *et al.*, 2022] Marco Godi, Christian Joppi, Geri Skenderi, and Marco Cristani. Movingfashion: a benchmark for the video-to-shop challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1678–1686, 2022.

[Gou *et al.*, 2023] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

[Guan *et al.*, 2023] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati. Leveraging pretrained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.

[Hazra *et al.*, 2024] Rishi Hazra, Pedro Zuidberg Dos Martires, and Luc De Raedt. Saycanpay: Heuristic planning with large language models using learnable domain knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20123–20133, 2024.

[Huang *et al.*, 2022] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022.

[Huang *et al.*, 2024] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Leake *et al.*, 2017] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.*, 36(4):130–1, 2017.

[Li *et al.*, 2023] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.

[Lin *et al.*, 2021] Qin Lin, Nuo Pang, and Zhiying Hong. Automated multi-modal video editing for ads video. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4823–4827, 2021.

[Liu and Yu, 2023] Chang Liu and Han Yu. Ai-empowered persuasive video generation: A survey. *ACM Computing Surveys*, 55(13s):1–31, 2023.

[Liu *et al.*, 2019] Chang Liu, Yi Dong, Han Yu, Zhiqi Shen, Zhanning Gao, Pan Wang, Changgong Zhang, Peiran Ren, Xuansong Xie, Lizhen Cui, et al. Generating persuasive visual storylines for promotional videos. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 901–910, 2019.

[Liu *et al.*, 2021] Chang Liu, Han Yu, Zhiqi Shen, Ian Dixon, Yingxue Yu, Zhanning Gao, Pan Wang, Peiran Ren, Xuansong Xie, Lizhen Cui, et al. Enhancing viewing experience of generated visual storylines for promotional videos. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[Maaz *et al.*, 2023] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[Madaan *et al.*, 2024] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[Press *et al.*, 2022] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[Scott Armstrong, 2011] J Scott Armstrong. Evidence-based advertising: An application to persuasion. *International Journal of Advertising*, 30(5):743–767, 2011.

[Shinn *et al.*, 2024] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[Smith, 2006] Tim J Smith. *An attentional theory of continuity editing*. University of Edinburgh. College of Science and Engineering. School of . . . , 2006.

[Sun *et al.*, 2010] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2432–2439. IEEE, 2010.

[Tang *et al.*, 2022] Yunlong Tang, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. Multi-modal segment assemblage network for ad video editing with importance-coherence reward. In *Proceedings of the Asian Conference on Computer Vision*, pages 3519–3535, 2022.

[Tao *et al.*, 2020] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation learning using inter-intra contrastive framework. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2193–2201, 2020.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Valentini *et al.*, 2018] Chiara Valentini, Stefania Romenti, Grazia Murtarelli, and Marta Pizzetti. Digital visual engagement: influencing purchase intentions on instagram. *Journal of communication management*, 22(4):362–381, 2018.

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

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Yao *et al.*, 2024] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[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.

[Zhang *et al.*, 2024] Kun Zhang, Jiali Zeng, Fandong Meng, Yuanzhuo Wang, Shiqi Sun, Long Bai, Huawei Shen, and Jie Zhou. Tree-of-reasoning question decomposition for complex question answering with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19560–19568, 2024.