Weakly-Supervised Movie Trailer Generation Driven by Multi-Modal Semantic Consistency

Sidan Zhu[1] , **Yutong Wang**[1] , **Hongteng Xu**[2] and **Dixin Luo**[1,3,†]
[1]Beijing Institute of Technology, Beijing
[2]Renmin University of China, Beijing
[3]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai
sidan_zhu@bit.edu.cn, yutongwang1012@gmail.com, hongtengxu@ruc.edu.cn, dxluo611@gmail.com

Abstract

As an essential movie promotional tool, trailers are designed to capture the audience's interest through the skillful editing of key movie shots. Although some attempts have been made for automatic trailer generation, existing methods often rely on predefined rules or manual fine-grained annotations and fail to fully leverage the multi-modal information of movies, resulting in unsatisfactory trailer generation results. In this study, we introduce a weakly-supervised trailer generation method driven by multi-modal semantic consistency. Specifically, we design a multi-modal trailer generation framework that selects and sorts key movie shots based on input music and movie metadata (e.g., category tags and plot keywords) and adds narration to the generated trailer based on movie subtitles. We utilize two pseudo-scores derived from the proposed framework as labels and thus train the model under a weakly-supervised learning paradigm, ensuring trailerness consistency for key shot selection and emotion consistency for key shot sorting, respectively. As a result, we can learn the proposed model solely based on movie-trailer pairs without any fine-grained annotations. Both objective experimental results and subjective user studies demonstrate the superior performance of our method over previous works. The code is available at https://github.com/Dixin-Lab/MMSC.

1 Introduction

Trailers serve as a powerful tool to showcase a movie's highlights and stimulate audience interest. Unlike video summarization [Narasimhan *et al.*, 2021; Wang *et al.*, 2023; Narasimhan *et al.*, 2022], which selects video clips in their original chronological sequence in the video to provide an overview, trailer generation requires not only the selection of key movie shots but also their rearrangement to maintain suspense and avoid revealing the storyline [Thompson, 1999; Hauge, 2017]. At the same time, a well-produced trailer should ensure *multi-modal semantic consistency* for its visual, textual, and acoustic information. For example, when a movie shot depicts a heavy storm, the music often features dense drum beats, and the narration may vividly describe the disaster, together creating a tense atmosphere. Therefore, generating a high-quality trailer requires the expertise of the editors and is time-consuming and expensive.

Currently, various efforts have been made to generate movie trailers automatically. Existing trailer generation methods either summarize empirical movie shot editing patterns and rules from official trailers or require manually defined fine-grained labels to learn a trailer generation model in a supervised way. However, due to the diversity and complexity of trailers and the scarcity of large-scale labeled movie-trailer datasets, these methods are highly prone to overfitting and often suffer from poor generalization performance. In addition, most existing methods rely solely on single-modality data, such as muted movies and trailers (visual modality) or movie metadata (textual modality), thereby failing to satisfy the requirements of real-world commercial applications. Although some methods try to integrate multi-modal data, they fail to sufficiently explore the semantic consistency across multiple modalities within high-quality movie trailers. As a result, the trailers generated by the existing methods remain far from satisfactory.

To address the aforementioned problems, we propose **MMSC**, a novel automated trailer generation framework driven by **M**ulti-**M**odal **S**emantic **C**onsistency. As illustrated in Figure 1, given a movie with metadata and a piece of music, we formulate the trailer generation task as selecting and sorting key movie shots according to the music and the metadata (e.g., category tags and plot keywords of movies). A multi-modal model is designed to achieve this task and is trained in a weakly-supervised way on a collection of movie-trailer pairs. In particular, during training, the model first derives the trailerness pseudo-scores and emotion pseudo-scores for movie shots, based on the input movie and music. The trailerness pseudo-score measures the likelihood of each shot being selected for the trailer, while the emotion pseudo-score reflects the emotional intensity of each movie shot. Accordingly, we train the model under the supervision of the two pseudo-scores, encouraging the model to select and sort the movie shots that are semantically consistent with those in official trailers. In addition, given the generated trailer se-

† Corresponding Author.

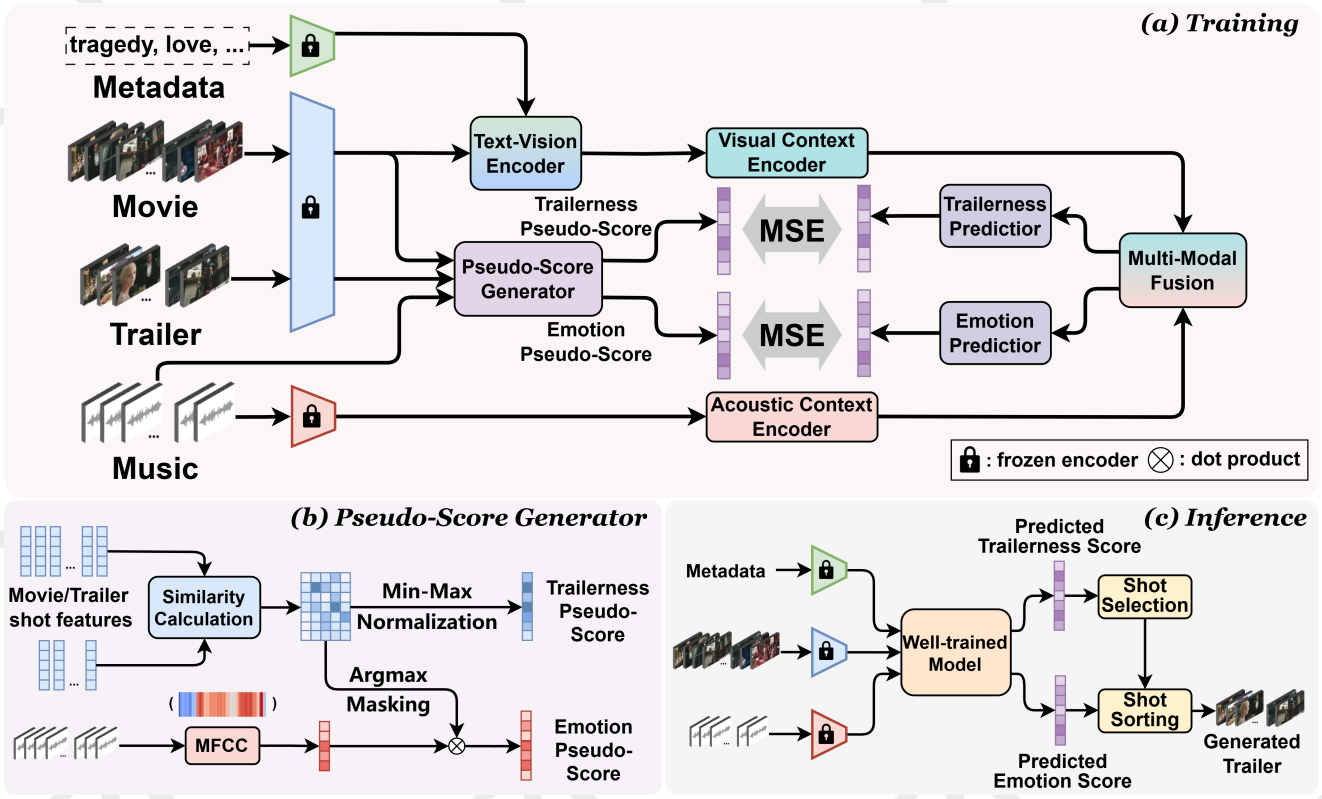

Figure 1: An illustration of our MMSC method. (a) Model architecture and loss components used during the training phase. (b) Calculation methods for trailerness and emotion pseudo-scores. (c) Trailer generation pipeline during the inference phase.

quence (i.e., the sorted key shots), we leverage a large language model [DeepSeek-AI *et al.*, 2024] to analyze movie subtitles and extract key sentences as the trailer narration, thereby improving the quality of the trailer.

In summary, our MMSC-based method utilizes multi-modal information to train a trailer generation model, ensuring semantic consistency across visual, textual, and acoustic modalities when generating trailers. Guided by the trailerness and emotion pseudo-scores, the proposed model generates semantically-meaningful trailers that effectively capture the styles and themes of the movies while aligning the emotion intensities with the given music. Moreover, our method is weakly supervised, learning from movie-trailer pairs without additional fine-grained annotations.

2 Related Work

Recently, many learning-based trailer generation methods have been proposed to achieve better performance. For example, some methods leverage movie subtitles [Hesham *et al.*, 2018] and movie plot summaries [Gaikwad *et al.*, 2021] as labels, selecting key movie shots that best match the labels. The work in [Papalampidi *et al.*, 2023] assigns each movie shot with a turning point label and sequentially selects the movie shots based on their labels when generating trailers. Besides using textual labels, some methods use frames [Liu and Jiang, 2015] or shots [Wang *et al.*, 2020] of official trailers as positive samples, training classifiers to identify useful frames or shots from input movies and stitching them as trailers. The work in [Argaw *et al.*, 2024] transforms the feature sequences of movies into those of trailers. More recently, some methods attempt to leverage multi-modal information for trailer generation. For example, the inverse partial optimal transport (IPOT) method in [Wang *et al.*, 2024] selects and sorts key movie shots via aligning the visual features of the movie shots with the acoustic features of music. The work in [Liu *et al.*, 2023] trains a trailer generation model using manually labeled emotional categories for each movie shot and selects key shots by maximizing the emotional alignment among visual, textual, and audio features.

However, the above learning-based methods require annotations for movies at the shot and even frame levels, which is expensive and time-consuming. The scarcity of such fine-grained labeled data makes these methods suffer a high risk of over-fitting. Although some attempts have been made to generate trailers using unsupervised learning methods, e.g., the anomaly detection-based movie-to-trailer (M2T) method in [Rehusevych, 2019], these methods often lead to sub-optimal performance due to the lack of annotations. In addition, most existing methods mainly rely on information from one or two modalities for trailer generation, and they seldom consider the multi-modal semantic consistency within high-quality trailers. Our work overcomes the limitations of the existing methods, achieving a weakly-supervised learning framework for trailer generation under the guidance of multi-

modal semantic consistency.

3 Proposed Method

Given a movie $\mathcal{M}$, a piece of music $\mathcal{A}$, and the movie's metadata $\mathcal{W}$, we aim to learn a multi-modal model to generate a trailer $\mathcal{T}$ based on the movie, the music, and the metadata jointly. Here, the trailer $\mathcal{T} = \{\mathcal{V}, \mathcal{A}\}$ consists of the video $\mathcal{V}$ extracted from the movie and the input music. We use the camera shot boundary detector TransNet-v2 [Soucek and Lokoc, 2024] to segment each movie and its corresponding trailer into shot sequences, i.e., $\mathcal{M} = \{m_i\}_{i=1}^{I}$ and $\mathcal{V} = \{v_j\}_{j=1}^{J}$, where m_i and v_j denote the visual shots of the movie and the trailer, respectively. For the music, we first apply the Ultimate Vocal Remover (UVR) tool[1] to remove its vocals and then segment it into music shots that are aligned with the trailer shots, denoted as $\mathcal{A} = \{a_j\}_{j=1}^{J}$.[2] The movie metadata consists of K tokens that indicate the movie's category labels and plot keywords, denoted as $\mathcal{W} = \{w_k\}_{k=1}^{K}$.

As illustrated in Figure 1, we design a multi-modal model, which takes $\{\mathcal{M}, \mathcal{A}, \mathcal{W}\}$ as input and generates $\mathcal{V}$ via selecting and sorting key movie shots based on the multi-modal semantic features of the input. The model is learned in a weakly-supervised learning framework, which does not require any fine-grained annotations of movies.

3.1 Model Architecture

As shown in Figure 1(a), our proposed model consists of three parts, i.e., a multi-modal encoder for data representation and two predictors for shot selection and sorting.

Multi-modal Encoding

We first apply the pre-trained ImageBind [Girdhar *et al.*, 2023] to derive the initial visual and textual features, respectively, i.e., $\boldsymbol{M} = f_v(\mathcal{M}) = [\boldsymbol{m}_i] \in \mathbb{R}^{I\times D}$, $\boldsymbol{V} = f_v(\mathcal{V}) = [\boldsymbol{v}_j] \in \mathbb{R}^{J\times D}$, and $\boldsymbol{W} = f_t(\mathcal{W}) = [\boldsymbol{w}_k] \in \mathbb{R}^{K\times D}$. Note that, the movie metadata used in this study includes movie labels and movie plot keywords, which reflect the core themes and style of the movie and are important considerations in trailer design. Therefore, the text-vision encoder merges the information of the metadata into the visual representation of each movie shot, enhancing the semantics of the visual representation accordingly. In particular, given $\boldsymbol{M}$ and $\boldsymbol{W}$, we deploy a cross-attention module between the source and the query modalities at the very first layers of the encoder, i.e.,

$$\boldsymbol{M}^t = \boldsymbol{M} + \sigma\Big(\frac{(\boldsymbol{M}\boldsymbol{W}_1^m)(\boldsymbol{W}\boldsymbol{W}_1^w)^\top}{\sqrt{D}}\Big)\boldsymbol{W}\boldsymbol{W}_2^w, \tag{1}$$

where $\{\boldsymbol{W}_i^m, \boldsymbol{W}_i^w \in \mathbb{R}^{D\times D}\}_{i=1}^{2}$, $\sigma(\cdot)$ denotes the softmax operation, and $\boldsymbol{M}^t \in \mathbb{R}^{I\times D}$ is the metadata-dependent visual representations of movie shots.

Furthermore, taking the metadata-dependent visual representations as input, we apply one more self-attention layer to capture the temporal correlations among the movie shots, i.e.,

$$\boldsymbol{M}^s = \mathrm{MLP}_M\Big(\sigma\Big(\frac{(\boldsymbol{M}^t\boldsymbol{W}_2^m)(\boldsymbol{M}^t\boldsymbol{W}_3^m)^\top}{\sqrt{D}}\Big)\boldsymbol{M}^t\boldsymbol{W}_4^m\Big), \tag{2}$$

where $\{\boldsymbol{W}_i^m \in \mathbb{R}^{D\times D}\}_{i=2}^{4}$, $\mathrm{MLP}_M : \mathbb{R}^D \mapsto \mathbb{R}^d$ is a multi-layer perceptron (MLP), and $\boldsymbol{M}^s = [\boldsymbol{m}_i^s] \in \mathbb{R}^{I\times d}$ is the contextualized visual features.

The music shots can be processed similarly. In particular, we apply the pre-trained CLAP [Wu *et al.*, 2023] to derive the initial acoustic features for the music shots, i.e., $\boldsymbol{A} = f_a(\mathcal{A}) = [\boldsymbol{a}_j] \in \mathbb{R}^{J\times D'}$, and obtain its contextualized acoustic features by a self-attention module, i.e.,

$$\boldsymbol{A}^s = \mathrm{MLP}_A\Big(\sigma\Big(\frac{(\boldsymbol{A}\boldsymbol{W}_1^a)(\boldsymbol{A}\boldsymbol{W}_2^a)^\top}{\sqrt{D'}}\Big)\boldsymbol{A}\boldsymbol{W}_3^a\Big), \tag{3}$$

where $\{\boldsymbol{W}_i^a \in \mathbb{R}^{D'\times D'}\}_{i=1}^{3}$, $\mathrm{MLP}_A : \mathbb{R}^{D'} \mapsto \mathbb{R}^d$, and $\boldsymbol{A}^s = [\boldsymbol{a}_j^s] \in \mathbb{R}^{J\times d}$ denotes the contextualized acoustic features of music shots.

Predictors for Shot Selection and Sorting

As aforementioned, our model selects and sorts movie shots conditioned on the given music. Therefore, we need to consider the inter-modal interaction between different modalities. In particular, we apply a cross-attention module to fuse the contextualized visual and acoustic features and get the final movie shot representation as

$$\boldsymbol{M}^c = \boldsymbol{M}^s + \sigma\Big(\frac{(\boldsymbol{M}^s\boldsymbol{W}_5^m)(\boldsymbol{A}^s\boldsymbol{W}_4^a)^\top}{\sqrt{d}}\Big)\boldsymbol{A}^s\boldsymbol{W}_5^a, \tag{4}$$

where $\boldsymbol{W}_5^m, \{\boldsymbol{W}_i^a\}_{i=4}^{5} \in \mathbb{R}^{d\times d}$, and $\boldsymbol{M}^c$ is the final movie shot representations, which are aligned cross-modal features for key shot selection and emotional prediction.

Given the final movie shot representations $\boldsymbol{M}^c$, we use two MLPs to predict the normalized trailerness score and emotion score for each movie shot, respectively, i.e.,

$$\begin{aligned}\hat{\boldsymbol{t}} &= [\hat{t}_i] = \mathrm{Sigmoid}(\mathrm{MLP}_t(\boldsymbol{M}^c)) \in [0,1]^I,\\ \hat{\boldsymbol{e}} &= [\hat{e}_i] = \mathrm{Sigmoid}(\mathrm{MLP}_e(\boldsymbol{M}^c)) \in [0,1]^I.\end{aligned} \tag{5}$$

Each $\hat{t}_i$ indicates the predicted probability that i-th movie shot is in the trailer and $\hat{e}_i$ indicates the predicted emotional intensity of i-th movie shot. The two predicted scores will be used for selection and sorting during the inference stage.

3.2 Weakly-supervised Learning Framework

We design two informative signals to supervise the training of the model, ensuring the trailerness and emotion scores derived in Eq. (5) to guide the selection and sorting of movie shots effectively. As illustrated in Figure 1(b), given a movie $\mathcal{M} = \{m_i\}_{i=1}^{I}$ and its corresponding trailer $\mathcal{T} = \{\mathcal{V}, \mathcal{A}\}$, where $\mathcal{V} = \{v_j\}_{j=1}^{J}$ and $\mathcal{A} = \{a_j\}_{j=1}^{J}$, we calculate two pseudo-scores based on the initial features and treat them as the training labels of our model. Because the labels are derived from the input data themselves, we essentially train our model in a weakly-supervised learning framework.

Trailerness Pseudo-Score

For each movie, we assign a trailerness pseudo-score t_i to the i-th movie shot based on the similarity between the movie shot and the shots in the corresponding trailer. Specifically, we first utilize the cosine similarity to compare each movie

[1] https://github.com/Anjok07/ultimatevocalremovergui.

[2] Here, we can use existing music segmentation tools such as Ruptures [Truong *et al.*, 2020] to obtain the music shots.

shot feature $\boldsymbol{m}_i$ with all trailer shots $\boldsymbol{V} = [\boldsymbol{v}_j]$ belong to the same movie, i.e.,

$$s_{i,j} = \frac{\boldsymbol{m}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{m}_i\|\|\boldsymbol{v}_j\|}, \ \forall\, i = 1, ..., I, \ j = 1, .., J. \tag{6}$$

Then, we take the maximum value in the set $\{s_{i,j}\}_{j=1}^{J}$ as the trailerness pseudo-score of the movie shot m_i, i.e.,

$$t_i = \max_{j \in \{1,...,J\}} s_{i,j}, \ \forall\, i = 1, 2, .., I. \tag{7}$$

A high trailerness pseudo-score t_i indicates that the movie shot m_i is highly similar to a trailer shot and should be selected, whereas a low pseudo-score suggests that it is dissimilar to any trailer shot and should not be included in the trailer.

Emotion Pseudo-Score

The video and the music of the official trailer are carefully designed to express emotions consistently [Xu *et al.*, 2015]. For example, music with a fast tempo is often paired with thrilling chase and fighting scenes, while slow and smooth music tends to complement scenes with a warm and soothing atmosphere. Therefore, besides the trailernesss pseudo-score, we further propose the emotion pseudo-score for each movie shot, quantifying its emotional intensity.

In particular, given a music shot, we calculate the average energy coefficient of its Mel Frequency Cepstral Coefficients (MFCCs), i.e., for each music shot a_j, we have

$$c_j = \text{Average}(\text{Norm}(\text{MFCC}(a_j))), \ \forall j = 1, 2, ..., J, \tag{8}$$

where MFCC$(\cdot)$ denotes extracting N_j energy coefficients, Norm$(\cdot)$ denotes min-max normalization, and Average$(\cdot)$ denotes the average operation. Note that, MFCC is commonly used to extract features from music, where the energy coefficient reflects core attributes of the music signal, such as timbre, intensity, frequency distribution, and dynamic variations. Music with high energy typically sounds loud and has a large dynamic range, a broad frequency spectrum, and strong rhythm and beats. Such music often conveys intense emotions and high vitality. Accordingly, a high emotion pseudo-score indicates an intense emotional expression in the music shot, while a low score suggests relatively mild emotions.

We transfer the emotion pseudo-scores from the music shots to the movie shots: given a movie shot, we first find the most similar trailer shot and then assign the emotion pseudo-score of the corresponding music shot to the movie shot, i.e., for $i = 1, 2, ..., I$,

$$e_i = c_j, \ \text{where } j = \arg\max_{j \in \{1,...,J\}} s_{i,j}. \tag{9}$$

Learning with Multi-modal Semantic Consistency

Given the above pseudo-scores and the predictions in Eq. (5), we learn the proposed model to fit the pseudo-scores, i.e.,

$$\mathcal{L}_{total} = \underbrace{\sum\nolimits_{i=1}^{I} |\hat{t}_i - t_i|^2}_{\mathcal{L}_t} + \underbrace{\sum\nolimits_{i=1}^{I} |\hat{e}_i - e_i|^2}_{\mathcal{L}_e}. \tag{10}$$

Here, the first loss pursues the trailerness consistency between a movie and its trailer based on their visual modality. The second loss pursues the emotion consistency between a movie and its trailer's music, which leverages the visual and acoustic modalities jointly. By minimizing the two losses jointly, we learn the proposed model, ensuring its prediction results to be consistent with the semantics within the corresponding trailers.

3.3 Trailer Generation

As illustrated in Figure 1(c), given a well-trained model, we can generate a trailer based on a movie, a piece of music, and the metadata of the movie through the following steps.

Selecting Key Movie Shots

First, given the movie shots $\boldsymbol{M} = [\boldsymbol{m}_i]_{i=1}^{I}$, the music shots $\boldsymbol{A} = [\boldsymbol{a}_j]_{j=1}^{J}$, and the metadata $\boldsymbol{W} = [\boldsymbol{w}_k]_{k=1}^{K}$, the model predicts the trailerness scores $\{\hat{t}_i\}_{i=1}^{I}$ and the emotion scores $\{\hat{e}_i\}_{i=1}^{I}$ for the movie shots. Accordingly, we select the movie shots corresponding to the J highest trailerness scores to construct the target trailer. The set of the movie shots, denoted as $\mathcal{M}_{\mathcal{S}} = \{m_i\}_{i \in \mathcal{S}}$, where $\mathcal{S} = \arg\text{Top-J}_{i \in \{1,...,I\}} \hat{t}_i$, covers the key shots that are likely to appear in the trailer.

Sorting Selected Movie Shots

As mentioned earlier, to maintain emotion consistency, the order of selected shots is crucial for constructing a high-quality trailer and should be aligned with the music shots. To achieve this aim, we sort the selected movie shots based on their emotion scores. In particular, given the music shots, we calculate their energy coefficients $\{c_j\}_{j=1}^{J}$ via Eq. (8), leading to a sequence reflecting the temporal dynamics of the emotions behind the music. We sort the selected shots based on their emotion scores $\{\hat{e}_i\}_{i \in \mathcal{S}}$, ensuring that the order of the sorted emotion scores matches with that of $\{c_j\}_{j=1}^{J}$.

Post-processing

Besides selecting and sorting movie shots, we further design three post-processing steps to make the generated trailer more realistic. Firstly, given the sorted key movie shots, we follow the previous work [Wang *et al.*, 2024] and align the duration of each shot with its corresponding music shot, as shown in Figure 2(a). For movie shots that exceed the duration of the corresponding music shots, we trim the excess parts of the movie shots. For movie shots that are shorter than the corresponding music shots, we use the adjacent shot from the original movie with higher trailerness score to fill the duration gap. Secondly, we use DeepSeek-V3 [DeepSeek-AI *et al.*, 2024], a pre-trained large language model (LLM), to analyze and select the movie's subtitles. As shown in Figure 2(b), the LLM takes the movie's subtitles with timestamps and some instructional prompts as input and selects some subtitles as the narrations of the generated trailer. Based on the timestamps of the subtitles, we can automatically extract the corresponding audio from the movie. Finally, we determine the positions of the selected narrations automatically. As shown in Figure 2(c), we utilize MiniCPM-V 2.6 [Yao *et al.*, 2024], a multi-modal LLM for video captioning, to generate a one-sentence description for each shot of the generated trailer. We extract the textual features of the shot descriptions and the selected narrations and calculate their pairwise similarities. Accordingly, we associate each narration with a shot by maximizing the sum of the similarities between all narrations and the shot descriptions under the constraint that the narrations do not overlap. This problem can be solved efficiently using dynamic programming (DP) [Bellman, 1966]. After adjusting the durations of the shots and merging the audio of the narrations into the music, we derive the final generated trailer.

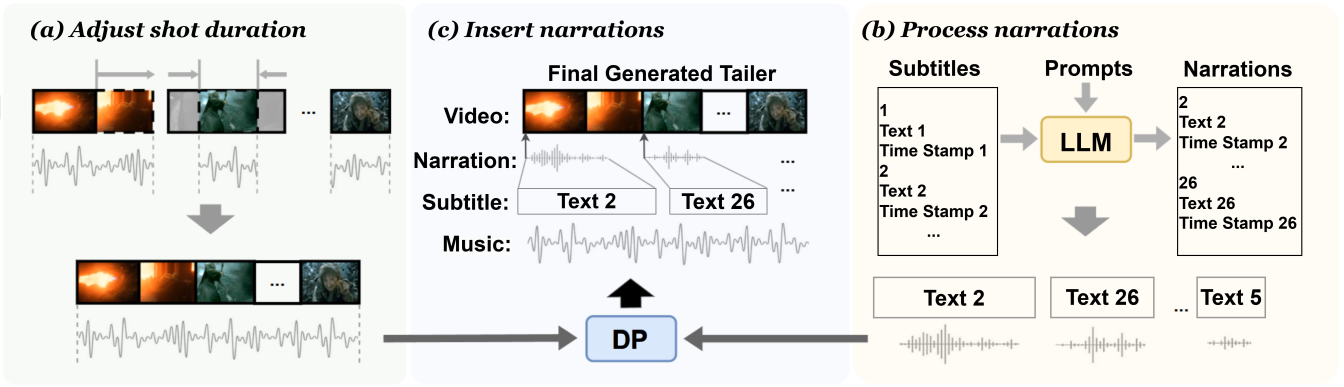

Figure 2: Illustration of post-processing steps, including adjusting the duration of selected movie shots (a), selecting narrations and processing narration audios (b), and inserting narrations into the trailer (c).

Category	Method	Test-8					Test-74				
		Precision↑	Recall↑	F1↑	LD↓	AA↑	Precision↑	Recall↑	F1↑	LD↓	AA↑
Video Summary	VASNet [Fajtl *et al.*, 2018]	0.0712	0.0645	0.0676	100.62	0.43	0.0496	0.0422	0.0455	84.79	0.44
	Muvee [Ganhör, 2014]	**0.2400**	0.0461	0.0714	103.50	0.36	–	–	–	–	–
	CLIP-It [Narasimhan *et al.*, 2021]	0.0711	0.0629	0.0667	101.12	0.45	0.0832	0.0710	0.0764	85.82	0.38
	OTVS [Wang *et al.*, 2023]	0.0688	0.0613	0.0648	101.25	0.46	0.0834	0.0711	0.0766	84.82	0.39
Trailer Generation	M2T [Rehusevych, 2019]	0.0611	0.0503	0.0515	**95.67**	0.42	–	–	–	–	–
	V2T [Irie *et al.*, 2010]	0.1121	0.0603	0.0945	103.75	0.52	–	–	–	–	–
	PPBVAM [Xu *et al.*, 2015]	0.0813	0.1244	0.0945	101.50	0.53	–	–	–	–	–
	TGT [Argaw *et al.*, 2024]	0.0703	0.0928	0.0788	124.75	0.47	0.0584	0.1001	0.0708	118.43	0.43
	IPOT (90%) [Wang *et al.*, 2024]	0.1187	0.1388	0.1277	102.75	0.44	0.1011	0.1184	0.1087	84.01	0.42
	IPOT [Wang *et al.*, 2024]	0.1218	0.1425	0.1311	101.50	0.42	0.1258	0.1483	0.1357	83.21	0.46
	MMSC (Ours)	0.1301	**0.1496**	**0.1391**	99.25	**0.58**	**0.1851**	**0.2163**	**0.1991**	**82.47**	**0.50**

Table 1: Experimental results compared with baselines on the two test sets. We bold the best results and underline the second-best results.

4 Experiments

4.1 Implementation Details

Dataset

We construct a dataset of 500 movies and 922 trailers spanning 18 genres and 30 years based on IMDb [IMDb, 2025] tags and years. The movie labels and movie plot keywords of each movie are collected from IMDb. Following the work in [Wang *et al.*, 2024], we resize movies in the dataset to 320p for learning convenience and efficiency and remove human vocals from the audio of trailers using UVR. The entire dataset is divided into training, validation, and test sets in a ratio of 85:5:10. The test set contains 74 movie-trailer pairs, denoted as **Test-74**. In addition, previous trailer generation methods [Ganhör, 2014; Rehusevych, 2019; Irie *et al.*, 2010; Xu *et al.*, 2015; Wang *et al.*, 2024] are tested on an independent test set with eight movies. Therefore, besides our dataset, we also use the eight movies as our test data, denoted as **Test-8**. The movies in Test-8 do not appear in the training and validation sets of our dataset.

Baselines

We compare our **MMSC**-based method with state-of-the-art trailer generation methods, including V2T [Irie *et al.*, 2010], M2T [Rehusevych, 2019], PPBVAM [Xu *et al.*, 2015], TGT [Argaw *et al.*, 2024], and IPOT [Wang *et al.*, 2024]. When evaluating the capability of shot selection, we also choose state-of-the-art video summarization methods as baselines, including a commercial video summarization software Muvee [Ganhör, 2014], VASNet [Fajtl *et al.*, 2018], CLIP-It [Narasimhan *et al.*, 2021], and OTVS [Wang *et al.*, 2023].

Evaluation Metrics

Following the work in [Wang *et al.*, 2024], we employ Precision, Recall, and F1-score (F1) to evaluate the performance of shot selection given the shots in the ground truth trailers (i.e., official trailers). Following the work in [Argaw *et al.*, 2024; Liu *et al.*, 2024], we employ Levenshtein distance (LD) and Pairwise agreement accuracy (AA) to evaluate the consistency between the order of the correctly selected movie shots with the order of ground truth trailer shots.

Model Architecture and Hyperparameter Settings

Each attention layer in our multi-modal model consists of two Transformer encoders with four attention heads. Each MLP in our model consists of two linear layers connected by a GELU function. The Adam [Kingma and Ba, 2015] optimizer is used during training, with an initial learning rate of 1e-5 and a cosine warm-up scheduler with β_1=0.9 and β_2=0.999. Our model is trained for 1,500 epochs with a batch size of 10, on a single NVIDIA RTX 3090.

Figure 3: Comparison between the trailer generated by our method against the official trailer. The gray shots in our generated trailer are incorrectly selected, while the rest are correct. The green checkmarks indicate that the movie shots connected by the double-headed arrows are in the correct order, while the red cross indicates an incorrect order. We also provide plot keywords and labels of the example movie, where plot keywords that are consistent with the shot contents are marked in blue, and shots that can reflect the movie labels are marked in purple. The last row shows the visualization of music energy coefficients, where a higher score is indicated by a darker red.

4.2 Objective Evaluation

In Table 1, we compare the performance of our method and baselines on movie shot selection and sorting. IPOT [Wang *et al.*, 2024] introduces a constraint in shot selection to avoid spoilers, that is, limiting the selection in the first 90% of the movie shots, i.e., IPOT (90%). However, to better evaluate the model's selection capability, we perform selection and ranking across all movie shots as other baselines. To ensure a fair comparison, we also provide the results of IPOT's selection across all movie shots in Table 1. TGT [Argaw *et al.*, 2024] uses the encoder-decoder architecture to generate the trailer sequence. Since there is no official trailer as input to limit the length of the trailer sequence generated by the TGT decoder during the inference phase, we set the length of the decoded sequence to no more than 150. From the results, we can see that our method achieves the best performance on most selection and sorting metrics. Especially on Test-74, our proposed method outperforms the state-of-the-art method IPOT by a large margin, i.e., 5.93%, 6.80%, and 6.34% on the Precision, Recall, and F1-score, respectively. Besides, the performance on two sorting metrics also improves by 5% and 4% in AA metric over the state-of-the-art method on the two test sets, respectively.

Since closely adjacent shots in a movie often share similar content, following prior work [Argaw *et al.*, 2024], we also relax the evaluation criteria by expanding the ground truth to include the R shots before and after each official trailer shot in the original movie. This prevents overly strict evaluation from overlooking reasonable shots and underestimating our model's performance. In Table 2, this expansion leads to improved model performance across all shot selection metrics, indicating our model's ability to generalize to shot variations.

Figure 3 provides an example comparing the trailer generated by our method with the official trailer. After selecting the appropriate movie shots for the trailer, our method calculates the energy coefficient for each shot of the given music, as visualized in the last row of Figure 3. In our generated trailer, shots with higher emotion scores correspond to the music shots with higher energy. This is attributed to emotion consistency loss, which adjusts the order of the selected movie shots to ensure that the emotion scores of the movie shots align with the energy coefficients of the music shots.

	Test-8			Test-74		
	Precision↑	Recall↑	F1↑	Precision↑	Recall↑	F1↑
$R=0$	0.1301	0.1496	0.1391	0.1851	0.2163	0.1991
$R=1$	0.2781	0.3208	0.2978	0.2772	0.3253	0.2985
$R=2$	**0.3803**	**0.4378**	**0.4069**	**0.3520**	**0.4126**	**0.3789**

Table 2: The analysis of relaxing the shot selection metrics. R represents the radius of the ground truth relaxation range.

4.3 Subjective Evaluation

Since the goal of producing trailers is to attract audiences, we conduct user studies to evaluate how well the trailers generated by our method and baselines perform compared to professionally edited trailers. We invite 30 volunteers (15 females and 15 males) to assess the quality of the trailers generated by different methods from four aspects: **Theme&Style**: "How well does the trailer convey the theme and style as the movie?", **Rhythm**: "How well do the visuals match the rhythm of the music?", **Attractiveness**: "How attractive is the trailer?", **Appropriateness**:"How close is the trailer to a real trailer?". Following the previous work [Xu *et al.*, 2015; Wang *et al.*, 2024], we resize all trailers to the same resolution, anonymize the names of the trailer generation methods, and upload them to the website for volunteers to rate. We also provide the official trailer (RT) as the reference standard with the highest score (set to seven). According to the average and median scores of the four aspects shown in Figure 4, our method significantly surpasses all baseline methods.

4.4 Ablation Study

Movie Metadata

We study the importance of metadata guidance by training our model with textual content at different granularity levels.

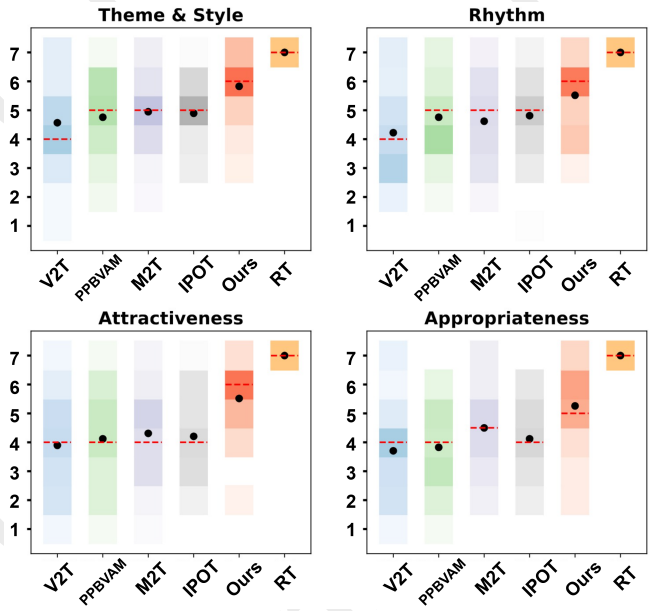

Figure 4: Comparison of subjective scores rated by volunteers to different methods. The black dots are means and the red dashed lines are medians. For each method, darker colors indicate a higher proportion of the corresponding score.

Label	Keyword	Test-8			Test-74		
		F1↑	LD↓	AA↑	F1↑	LD↓	AA↑
✗	✗	0.1218	99.65	0.42	0.1871	83.69	0.40
✓	✗	0.1334	99.62	0.48	0.1942	83.00	0.44
✗	✓	0.1360	99.50	0.54	0.1944	82.86	0.49
✓	✓	**0.1391**	**99.25**	**0.58**	**0.1991**	**82.47**	**0.50**

Table 3: Ablation studies on the impacts of movie metadata.

The movie labels represent the narrative style of the movie, providing coarse-grained categorical information. Movie plot keywords capture the core themes and key visual elements of the movie, providing fine-grained content descriptions. In Table 3, training the model without any metadata leads to severe performance degradation across all metrics, while applying metadata at both granularity levels contributes to model performance. Compared to abstract labels, the more specific plot keywords can better guide the model to learn key events that are more likely to be selected by professional editors.

Network Components

In Table 4, we analyze the impact of different network components by training the model with the single-modality context encoder or without any context encoders. The results show that both context encoders enhance overall model performance by integrating contextual information and improving feature discrimination between similar movie shots or similar music shots. Notably, the models corresponding to the suboptimal results differ between the two test sets. We speculate that this may result from Test-8, which has fewer trailers and is more sensitive to atypical trailers, leading to inconsistencies with the larger test set Test-74.

Visual Encoder	Acoustic Encoder	Test-8			Test-74		
		F1↑	LD↓	AA↑	F1↑	LD↓	AA↑
✗	✗	0.1289	99.75	0.51	0.1677	83.10	0.47
✗	✓	0.1380	99.50	0.55	0.1702	83.04	0.46
✓	✗	0.1318	99.50	0.52	0.1725	82.90	0.48
✓	✓	**0.1391**	**99.25**	**0.58**	**0.1991**	**82.47**	**0.50**

Table 4: Ablation studies on the impacts of network components. Visual Encoder and Acoustic Encoder refer to Visual Context Encoder and Acoustic Context Encoder for brevity.

Sorting Method	Test-8			Test-74		
	DC↑	AC↑	HM↑	DC↑	AC↑	HM↑
Random	0.38	0.46	0.38	0.45	0.41	0.39
Chronological	0.00	1.00	0.00	0.00	1.00	0.00
Proposed	0.52	0.52	**0.44**	0.51	0.47	**0.43**

Table 5: Ablation studies on the impacts of emotion consistency.

Emotion Consistency Loss

To investigate the effectiveness of emotion consistency loss in shot sorting, we compare shot sequences obtained by sorting randomly (i.e., Random), sorting according to the chronological order in the movie (i.e., Chronological), and sorting based on the emotion consistency (i.e., Proposed) against the official trailer sequence. We propose three metrics to measure the detailed order consistency, namely, Ascending consistency (AC), Descending consistency (DC), and the harmonic mean (HM) calculated based on these two metrics, i.e., $\text{HM} = \frac{2\times\text{AC}\times\text{DC}}{\text{AC}+\text{DC}}$. Note that for any pair of two different shots (m_i, m_j), where $i < j$, with their shot indices (a_i, a_j) in the generated trailer, there are two possible order relationships of these two shots in the ground truth trailer, including ascending order $a_i < a_j$ and descending order $a_i > a_j$. The AC metric computes the number of correctly predicted pairwise ascending orders in the generated trailer among all possible pairwise ascending orders. The DC metric does the same for descending orders. As shown in Table 5, since we calculate metrics on the shot sequence of correctly selected shots by our model and all pairwise orders in the chronological sorted sequence are in ascending order, DC and AC of the chronologically sorted sequence are always fixed at 0 and 1, respectively. Compared with the randomly sorted and the chronologically sorted sequences, the sequence generated based on emotion consistency achieves the best performance in DC and HM, demonstrating that our proposed strategy effectively captures reverse-order editing techniques used in official trailers to prevent spoilers and enhance suspense.

5 Conclusion

In this work, we propose a weakly-supervised framework for trailer generation, driven by multi-modal semantic consistency. Our method leverages data from all modalities to generate movie trailers that contain all modality elements. The entire process is fully automated, eliminating the need for additional fine-grained annotation. Experiments demonstrate that our MMSC-based method outperforms state-of-the-art trailer generation and video summarization methods on both objective and subjective evaluation metrics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62102031), and the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai, P.R. China (AI202409).

References

[Argaw *et al.*, 2024] Dawit Mureja Argaw, Mattia Soldan, Alejandro Pardo, Chen Zhao, Fabian Caba Heilbron, Joon Son Chung, and Bernard Ghanem. Towards automated movie trailer generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7445–7454, 2024.

[Bellman, 1966] Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

[DeepSeek-AI *et al.*, 2024] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and so on. Deepseek-v3 technical report, 2024.

[Fajtl *et al.*, 2018] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer, 2018.

[Gaikwad *et al.*, 2021] Bhagyashree Gaikwad, Ankita Sontakke, Manasi Patwardhan, Niranjan Pedanekar, and Shirish Karande. Plots to previews: Towards automatic movie preview retrieval using publicly available metadata. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3205–3214, 2021.

[Ganhör, 2014] Roman Ganhör. Muvee: An alternative approach to mobile video trimming. In *IEEE International Symposium on Multimedia*, pages 229–236. IEEE, 2014.

[Girdhar *et al.*, 2023] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[Hauge, 2017] Michael Hauge. *Storytelling Made Easy: Persuade and Transform Your Audiences, Buyers, And Clients-Simply, Quickly, and Profitably*. BookBaby, 2017.

[Hesham *et al.*, 2018] Mohammad Hesham, Bishoy Hani, Nour Fouad, and Eslam Amer. Smart trailer: Automatic generation of movie trailer using only subtitles. In *2018 First International Workshop on Deep and Representation Learning (IWDRL)*, pages 26–30. IEEE, 2018.

[IMDb, 2025] IMDb. https://www.imdb.com. 2025.

[Irie *et al.*, 2010] Go Irie, Takashi Satou, Akira Kojima, Toshihiko Yamasaki, and Kiyoharu Aizawa. Automatic trailer generation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 839–842, 2010.

[Kingma and Ba, 2015] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Liu and Jiang, 2015] Xingchen Liu and Jianming Jiang. Semi-supervised learning towards computerized generation of movie trailers. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 2990–2995. IEEE, 2015.

[Liu *et al.*, 2023] Wu-Qin Liu, Min-Xuan Lin, Hai-Bin Huang, Chong-Yang Ma, Yu Song, Wei-Ming Dong, and Chang-Sheng Xu. Emotion-aware music driven movie montage. *Journal of Computer Science and Technology*, 38(3):540–553, 2023.

[Liu *et al.*, 2024] Meng Liu, Mingda Zhang, Jialu Liu, Hanjun Dai, Ming-Hsuan Yang, Shuiwang Ji, Zheyun Feng, and Boqing Gong. Video timeline modeling for news story understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[Narasimhan *et al.*, 2021] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021.

[Narasimhan *et al.*, 2022] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[Papalampidi *et al.*, 2023] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Finding the right moment: Human-assisted trailer creation via task composition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Rehusevych, 2019] Orest Rehusevych. movie2trailer: Unsupervised trailer generation using anomaly detection. 2019.

[Soucek and Lokoc, 2024] Tomás Soucek and Jakub Lokoc. 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, 2024.

[Thompson, 1999] Kristin Thompson. *Storytelling in the new Hollywood: Understanding classical narrative technique*. Harvard University Press, 1999.

[Truong *et al.*, 2020] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[Wang *et al.*, 2020] Lezi Wang, Dong Liu, Rohit Puri, and Dimitris N Metaxas. Learning trailer moments in full-length movies with co-contrastive attention. In *European Conference on Computer Vision*, pages 300–316. Springer, 2020.

[Wang *et al.*, 2023] Yutong Wang, Hongteng Xu, and Dixin Luo. Self-supervised video summarization guided by semantic inverse optimal transport. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6611–6622, 2023.

[Wang *et al.*, 2024] Yutong Wang, Sidan Zhu, Hongteng Xu, and Dixin Luo. An inverse partial optimal transport framework for music-guided trailer generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9739–9748, 2024.

[Wu *et al.*, 2023] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Xu *et al.*, 2015] Hongteng Xu, Yi Zhen, and Hongyuan Zha. Trailer generation via a point process-based visual attractiveness model. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[Yao *et al.*, 2024] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.