MoviePuzzle: Visual Narrative Reasoning through Multimodal Order Learning

Jianghui Wang[1,*], Yuxuan Wang[2,*], Dongyan Zhao[2], Zilong Zheng[1,✉]
[1] Beijing Institute for General Artificial Intelligence
[2] Wangxuan Institute of Computer Technology, Peking University
uangjianghui@gmail.com, {wyx,zhaody}@pku.edu.cn, zlzheng@bigai.ai
https://moviepuzzle.github.io

Abstract

We introduce **MoviePuzzle**, *a novel challenge that targets visual narrative reasoning and holistic movie understanding. Despite the notable progress that has been witnessed in the realm of video understanding, most prior works fail to present tasks and models to address* ***holistic video understanding*** *and the innate* ***visual narrative structures*** *existing in long-form videos. To tackle this quandary, we put forth* **MoviePuzzle** *task that amplifies the temporal feature learning and structure learning of video models by reshuffling the shot, frame, and clip layers of movie segments in the presence of video-dialogue information. We start by establishing a carefully refined dataset based on MovieNet [25] by dissecting movies into hierarchical layers and randomly permuting the orders. Besides benchmarking the* **MoviePuzzle** *with prior arts on movie understanding, we devise a Hierarchical Contrastive Movie Clustering (*HCMC*) model that considers the underlying structure and visual semantic orders for movie reordering. Specifically, through a pairwise and contrastive learning approach, we train models to predict the correct order of each layer. This equips them with the knack for deciphering the visual narrative structure of movies and handling the disorder lurking in video data. Experiments show that our approach outperforms existing state-of-the-art methods on the* **MoviePuzzle** *benchmark, underscoring its efficacy. All the datasets and baseline codes will be publicly accessible.*

1. Introduction

Humans, even young kids, are capable of quickly perceiving and comprehending different forms of visual media, such as comics, short videos, 3D movies, *etc*. Without paying attention to details, we **connect** key visual or auditory information and **reason** over them in real-time to form a summarized *visual narrative* [14]. Consider movie frames in Figure 1, one can quickly understand its story as: *the male is scamming money over the phone*[1], without counting the accurate number of people in panel **E**. As such, we can watch and understand long-form videos, movies, and tens of episodes of TV shows.

Figure 1. A typical example of **MoviePuzzle**. Five frames in a clip from the movie *The Wolf of Wall Street* are randomly shuffled and placed. You are asked to connect the pieces with rational logic to form a plausible narrative. The letters marked on the upper-left corners are reference indexes and some (not all) frames are tagged with subtitles. The ground truth order is left in the footnote of this page.

As for the computer vision community, video understanding has achieved significant improvements over the past few years, with various benchmarks proposed on fine-grained action predictions [64, 17], Video Question Answering (VideoQA) [19, 36, 54], and video-grounded dialogue generations [32, 31], *etc*. The *de facto* paradigm is to integrate all extracted dense features of vision and language into a fusion layer w.r.t. their temporal orders [60, 4]. Despite the inspiring progress of Video-Language models on

*Equal contribution. Correspondence to Zilong Zheng <zlzheng@bigai.ai>.

[1] The correct order is $A \rightarrow B \rightarrow D \rightarrow C \rightarrow E$.

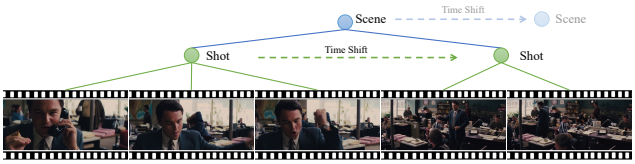


Figure 2. Illustration of movie visual narrative structures. The leaf nodes are video frames connected with temporal consistency on action dynamics and appearance information; the middle-level nodes are shot-level summaries connected with event consistency; the top-level nodes are scene-level synopses connected with narrative consistency.

these tasks, the benchmarks and models commonly neglect two key factors:

Holistic Video Understanding: Fundamentally, most modern literature deviates from the natural setting of comprehending videos in a holistic and abstract manner, *i.e.*, the tasks are treated as multi-frame image understanding problems that target inner-frame and inter-frame modeling. The linearly growing computational cost largely limits the potential of consuming long-form videos like movies and TV shows, which are fulfilled with scene and shot transitions. It is worth noting that some recent works [25] have observed the significance of holistic understanding, while the tasks are still defined as visual recognition and classifications.

Visual Narrative Structures: Beyond the individual visual surface and subtitle lexicons, the ability to perceive and understand a sequence of multimodal images is closely related to the underlying *narrative structure*, a hierarchical understanding that guides the presentation of events and individual concepts [13]. Figure 2 depicts an exemplar illustration: each frame is considered as a brick to form the building of the movie narrative, where any irrational swap between frames may break the integrity all of a sudden.

To address the deficiencies of the prior work and step towards holistic long-form video understanding, we present a new challenge **MoviePuzzle** that aims at benchmarking the machine's capabilities on visual narrative reasoning (VNR). In this task, the machine is asked to reorder the sequences of frames to form a plausible narrative; see Figure 1 for an example. We establish a carefully refined dataset based on MoiveNet [25]: all frames and corresponding subtitles in a clip are well aligned to form a human-understandable narrative. We further limit the maximum number of the scene and shot transitions within a clip to lower the computational expense of visual perception. The underlying rationale that we choose frame reordering as the main task derives from the preliminary diagnostic studies on visual narrative structures [13]. It is worth noting that some works have considered visual reordering as one of the downstream applications [67, 52, 16], yet the experiments are mainly on local temporal consistency or procedural understanding.

Solving **MoviePuzzle** is intrinsically non-trivial compared with single-frame understanding tasks. We highlight the external requirements on models brought out by VNR: (i) *Commonsense Reasoning*: the basic knowledge of visual commonsense (underlying rationale of visual activities [66]) and social commonsense (emotional and social activity understanding [50]); (ii) *Visual-dialogue Grounding*: the ability to integrate information within weakly-aligned dialogues and visual frames; (iii) *Visual-dialogue Summarization*: the ability to connect individual frames to form a narrative. To probe the machine's capability for narrative summarization, we add a downstream application as movie synopsis association, where the model is to extract the corresponding synopsis piece with pre-trained VNR models. Besides, we test the generalizability of neural models by splitting the test set into the in-domain test and out-domain test, where the latter are clips selected from unseen movies only with their pre-extracted visual features. We benchmark **MoviePuzzle** with prior successful models on Movie understanding. To address the structured information existing in the narrative, we further devise a contrastive-learning (CL)-based hierarchical representation for VNR.

To summarize, our contributions are three-fold: (i) We introduce **MoviePuzzle**, a novel task that aims to facilitate the learning of underlying temporal connections between different plot points in movies by reassembling multimodal video clips; (ii) We construct **MoviePuzzle** dataset that aligns video clip images and subtitles on a sentence-by-sentence basis. Moreover, we provide annotations for the structural hierarchy of frames, shots, and scenes, all of which are aligned at their boundaries; (iii) We devise a new Hierarchical Contrastive Movie Clustering (HCMC) model and establish a benchmark. Our model outperforms the current baseline models and achieves state-of-the-art results.

2. The MoviePuzzle Task

The **MoviePuzzle** contains a diverse range of data from various modalities and rich annotations on different aspects of movie structures, enabling a comprehensive understanding of movie content. Figure 1 showcases an exemplar datapoint available for the movie *The Wolf of Wall Street*.

2.1. Dataset Curation

Movie We extract movie sources from the existing large-scale movie dataset pool, MovieNet [25]. Although MovieNet has data sources from 1,100 movies, most of the data are coarse. First, not all movies have the required labels or semantic information. Second, the dataset does not align subtitle dialogues with each frame. Third, it does not slice the movies into semantically reasonable segments. While such rough annotations are acceptable for the benchmark provided by MovieNet, our task of fine-grained reordering of a multimodal movie hierarchical structure requires frame-level annotations. We carefully filter movies from

the original pool by ensuring the existence of frame image, corresponding subtitle file, frame-to-shot boundary, frame-to-scene boundary, cast id, and synopsis, resulting in 228 movies with rich tags and a complete semantic structure.

Image & Subtitle We collect 135,837 images from the metadata. All images are extracted as RGB images from 240P movies using screenshots, and cropped to remove black borders. In terms of subtitles, we select English subtitle files that are aligned with the movie version. Movies with few dialogues will be discarded. For all subtitles, we discard symbols that are not in Unicode, use the Latin alphabet to represent letters with accents (foreign words) uniformly, and filter out onomatopoeic subtitles and narrations that represent environmental sounds. Figure 5 shows the wordle of all utterances. As seen, most lexicons have few relationships with visual objects and scenes, yielding extra challenges to vison-dialogue alignment and understanding.

2.2. Refined Data Annotation

Image-text pairs The original MovieNet dataset does not consistently provide corresponding images for each dialogue. Here, we pair each frame image with its corresponding text, achieving the finest granularity alignment of images and text on the time sequence. Each dialogue is only matched with the images falling exactly within its corresponding time period, rather than the nearest ones, ensuring the maximum matching of image and text content. In the case of multiple frames falling on a single dialogue, we choose to merge these frames and select the middle frame as the most representative matching image. For most of the original images that do not match with subtitles, we discard them unless they are exceptional examples within a clip for an extended period. Finally, our image-subtitle pairs account for 83.5% of the total frames.

Scene&Shot aligned boundary Regarding the visual narrative structure of movies as in Figure 2, most clips have two innate hierarchical levels: shots and scenes. A shot is a fixed camera angle capture and typically a short sequence of temporally consecutive frames indicating a few actions. A scene is a sequence of shots that share the same environmental context, which typically depicts an event or a short story. Capturing the hierarchical structure of a movie is vital for movie understanding. Of note, scene segmentation remains an open problem to video understanding [9, 48, 47, 10]. We leverage annotations from MovieNet [25] to form our labels, including the manually annotated scene boundaries and automatically generated shot boundaries using [51]. In total, the MoviePuzzle has 15,414 scenes and over 81K semantically rich shots.

Clip aligned boundary In order to obtain video clips suitable for training, we further segment the 228 movies and ultimately obtain 10,031 movie clips. We follow the following criteria to select clips: First, each clip consists of 10 to 20 frames, and the length of the clip is controlled to tell a short plot without the semantic information being too monotonous; Second, the image-text pairs in each clip must be greater than 80%, ensuring the semantics are not too sparse. As shown in Figures 3 and 4, most clips have 1 or 2 scenes (88%) and 5 to 10 separated shots (75.91%), which is in line with the structure of a short plot.

2.3. Data Statistics

We divide the entire dataset into five parts: train, val, in-domain test, and out-domain test (Table 1), each accounting for 70%, 6%, 12%, and 12% of the total clips, respectively. The data in val and in-domain test all come from the same set of movies as the train split. We select val and in-domain test by taking equally spaced clips based on the clip numbers from the movie clips covered by the train dataset. The out-domain test split is entirely taken from different movies as in the train split. This division helps assess the model's generalization ability.

	all	train	val	in-domain test	out-domain test
#. of Clips	10,031	7,048	589	1,178	1,196

Table 1. Statistics of MoviePuzzle Splits.

3. Benchmarking the MoviePuzzle

3.1. Task Formulation

MoviePuzzle consists of a set of boundary aligned image-text clips $(V, U, A, B) \in \mathcal{D}$. Here V, U, A, B are ordered sets have same length N denoted the frame numbers of the movie clip. $V = \langle v_1, v_2, \cdots, v_N \rangle$ serves as an image clip with v_i denoting the i-th input frame. $U = \langle u_1, u_2, \cdots, u_N \rangle$ signifying the corresponding subtitles. $A = \langle a_1, a_2, \cdots, a_N \rangle$ representing the shot label list where $a_i \in [1, N^{shot}]$ is the i-th frame's shot id, N^{shot} is the shot numbers of the movie clip. $B = \langle b_1, b_2, \cdots, b_N \rangle$ is the scene label list with $b_i \in [1, N^{scene}]$ as the i-th frame's scene id, , N^{scene} is the scene numbers of the movie clip. In the given movie clip (V, U, A, B), the (v_i, u_i, a_i, b_i) represents the information encapsulated within a single frame. It is imperative to note that the subscript i denotes the index of the frames after they have been subjected to a same randomized reordering process.

In our tasks, the $GT = \langle l_1, l_2, \cdots, l_N \rangle$ indicates the correct ordering sequence of the video clips. Our goal is to predict a possible index sequence $Pred = \langle \hat{l}_1, \hat{l}_2, \cdots, \hat{l}_N \rangle$ for image-text clips that are as close as possible to the true order GT.

3.2. Evaluation Metric

Measuring the validity of a predicted order is non-trivial. First, since we only present sliced frame pieces and not all

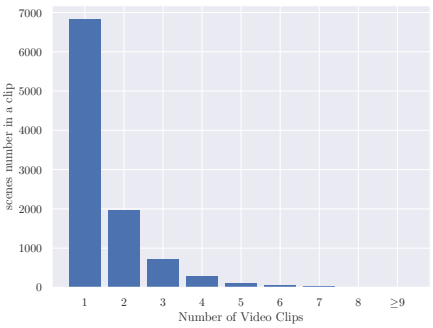


Figure 3. Distribution of scenes in a clip.

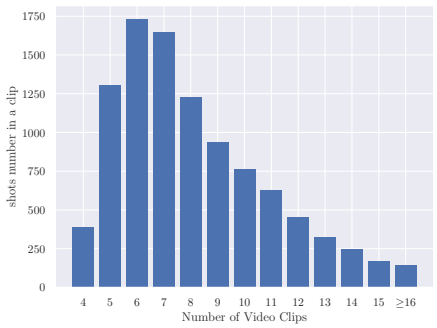


Figure 4. Distribution of shots in a clip.

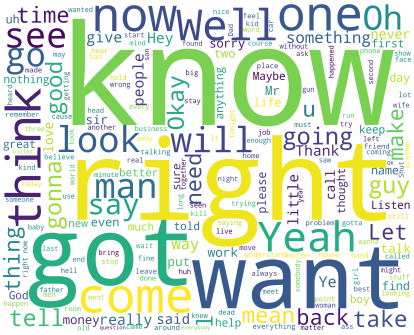


Figure 5. Wordcloud of utterances.

of them are tagged with semantic subtitles, there might be multiple ways to form a plausible visual narrative [14]. For example, the frame sequences of sunset and sunrise shall both be considered rational, even though the order is entirely reversed. Second, there is no unified answer to measure the matching score between the predicted order and the ground truth. Below we derive several ways to evaluate the ordering task's metrics.

Pairwise Score The pairwise matching score is one of the most commonly used metrics for permutations in discrete mathematics and permutation graph theory. Suppose the video clip has N frames, we denote the ground truth sequence as $GT = \langle l_1, l_2, \cdots, l_N \rangle$, and model ordering prediction sequence as $Pred = \langle \hat{l}_1, \hat{l}_2, \cdots, \hat{l}_N \rangle$, then a match is a pair of elements that follow their natural order: $i < j$ and $l_i \prec l_j$ and $\hat{l}_i \prec \hat{l}_j$, where $\prec$ is a comparator indicating the precedent temporal order. The pairwise score is computed as the ratio of the cardinality of the match set:

$$PairScore = \frac{\#\{(i,j) | i < j \wedge l_i \prec l_j \wedge \hat{l}_i \prec \hat{l}_j\}}{\binom{N}{2}}, \quad (1)$$

where $\binom{N}{2} = n(n-1)/2$ is the total number of index combinations.

Triplet Score The pairwise score can be easily generalized to measure the temporal orders across three frames. A triplet match can be defined as $i < j < k$ and $l_i \prec l_j \prec l_k$ and $\hat{l}_i \prec \hat{l}_j \prec \hat{l}_k$. The triplet score shares a similar equation as in Eqn. (1) but changes the pairwise match to triplets.

3.3. Hierarchical Contrastive Movie Clustering

3.3.1 Hierarchical Movie Representations

Frame Level Representation The movie clip (V, U, A, B) firstly passes through an encoder $E : (v, u) \rightarrow (v', u')$ to compute the image and text feature tokens, respectively. We aggregate image and text feature tokens by concatenating them across the second dimension *i.e.* sequence length to compute frame level representation $X = \langle x_1, x_2, \cdots, x_N \rangle$ where $x_k = v_k{}' \oplus u_k{}'$ indicating the frame k. Concatenation along the second dimension allows us to combine the two modalities, $x_k \in \mathbb{R}^{c \times d}$, where c is the total tokens length and d is the size of the multimodal hidden states.

Shot Level Representation Based on the shot label A, we divide X into N_{shot} shots groups to gain shot level representation $F = \langle f_1, f_2, \cdots, f_{N^{shot}} \rangle$ with the same group labels. Each group contains frames with the same shot label. For a shot clip list $f_k \in F$, we denote $f_k = \langle x_{k1}, x_{k2}, \cdots, x_{kn} \rangle$ as shot k representation with each $x_{ki} \in f_k$ denoting a frame's representation.

In order to forecast the temporal relationship between any two frames $(x_{ki}, x_{kj}) \in f_k$ within a single shot k, we concatenate them together as $f_{ij}^{*k} = \phi_{frame}(x_{ki}) \oplus \phi_{frame}(x_{kj})$, with the left frame i representing a preceding temporal position to the right frame j, and ϕ_{frame} is the frame level learnable embedding layer. Thus $F^{*k} = \langle f_{ij}^{*k} \rangle$ contains all the information needed to rank shot k.

Scene Level Representation Based on the scene scene label B, we divide X into N^{scene} shot groups $G = \langle g_1, g_2, \cdots, g_{N^{scene}} \rangle$ with the same scene group labels. Each group contains frames with the same scene label. Then according to the shot label A, we can further divide $g_k \in G$ into N_k^{shot} shot groups $g_k = \langle f_1, f_2, \cdots, f_{N_k^{shot}} \rangle$. g_k represents the kth scene layer shots, each $f_{ki} \in g_k, f_{ki} = \langle x_{ki1}, x_{ki2}, \cdots, x_{kin} \rangle$ is a shot clip. We use $\overline{f}_{ki} = x_{ki1} \oplus \cdots \oplus x_{kin}$ to indicate the representation of one shot.

Any two shots f_{ki}, f_{kj} can be concatenated together as $g_{ij}^{*k} = \phi_{shot}(\overline{f}_{ki}) \oplus \phi_{shot}(\overline{f}_{kj}))$, and ϕ_{shot} is the shot level learnable embedding layer, with the left shot representing a preceding temporal position to the right frame. Later, we will predict the internal ordering of any two shots with $G^{*k} = \langle g_{ij}^{*k} \rangle$.

Clip Level Representation The scene layer represents the top level of a clip. Based on the scene label B, we divide X into N^{scene} scene groups $H = \langle h_1, \cdots, h_{N^{scene}} \rangle$ with the same scene labels. For a scene clip list $h_k \in H$, we denote $h_k = \langle x_{k1}, \cdots, x_{kn} \rangle$ as scene k representation, for all $x_{ki} \in h_k$ denoting one frame's representation.

We use $\overline{h}_k = x_{k1} \oplus \cdots \oplus x_{kn}$ to indicate the representation of scene k. Any two scenes h_i, h_j can be concatenated together as $h^*_{ij} = \phi_{scene}(\overline{h}_i) \oplus \phi_{scene}(\overline{h}_j))$, with the left frame representing a preceding temporal position to the right frame. We use $H^* = \langle h^*_{ij} \rangle$ to predict the internal ordering of two scenes.

3.3.2 Modeling and Learning

An overview of the Hierarchical Contrastive Movie Clustering (HCMC) model is sketched in Figure 6. At the training stage, we jointly train our model in an efficient end-to-end way. Specifically, the input of our model is three levels of video-dialogue pairs, including single frame, shot sets, and scene sets. We jointly optimize our model with three-level ordering tasks and two-level clustering tasks.

Multimodal Feature Extractor Since most popular video-language works [67, 52, 16] are not able to obtain the long dependency temporal information, we adopt the end-to-end way to train our model, aiming to learn a better feature extractor. To mitigate the gap between vision and language, we initialize the vision encoder and language encoder with the parameters of CLIP [46] vision encoder and language encoder, respectively.

Video-Dialogue Transformer Encoder After obtaining the embedding of frame and dialogue, we jointly optimize our model with a temporal ordering learning task and a contrastive learning task. For the ordering task, we first concatenate two randomly sampled image-dialogue pairs in the same level and then feed them to the frame-dialogue transformer encoder, after that we use a binary classification head to learn if the input is in the true temporal order. It is worth noting that to mitigate the exposure error, we sample the frame-level pairs not constrained to ones belonging to the same shot. For the clustering task, we randomly sampled image-dialogue pairs from different groups as negative samples, we then optimize the cluster head by pulling together the positive sample and pushing away the negative samples.

Objective We randomly sample an ordered pair of the same layer clips $(\mathcal{P}_i, \mathcal{P}_j, \widetilde{\mathcal{Q}})$ from the extracted embeddings, where $\mathcal{P}_i$ and $\mathcal{P}_j$ are two random embeddings belonging to the same group and $\widetilde{\mathcal{Q}} = \{\widetilde{\mathcal{Q}}_1, \widetilde{\mathcal{Q}}_2, \cdots, \widetilde{\mathcal{Q}}_n\}$, where n is the number of negative samples, is the negative data sampled outside the current permutation group. The pair-wise layer representation $(\mathcal{P}_i, \mathcal{P}_j)$ is fed into a binary classifier ϕ, with two classes indicating the forward order $(\mathcal{P}_i \succ \mathcal{P}_j)$ and the backward order $(\mathcal{P}_i \prec \mathcal{P}_j)$ of input pairs. The cross-entropy loss is calculated based on the classifier output. Formally, the objective can be denoted as:

$$\mathcal{L}_{cls}(\mathcal{P}_i, \mathcal{P}_j) = CrossEntropy(\phi(\mathcal{P}_i, \mathcal{P}_j), O_{GT}), \quad (2)$$

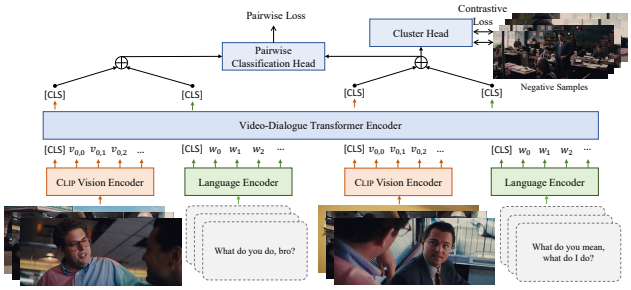


Figure 6. An overview of frame-level HCMC model.

where $O_{GT} \in \{0, 1\}$ indicates the ground-truth ordering class.

To derive a shot and scene-aware representation, we leverage the Contrastive Learning trick by feeding the output representation to another multi-class classifier ψ, where positive and negative data are distinguished based on the contrastive objective:

$$\mathcal{L}_{CL}(\mathcal{P}_i, \mathcal{P}_j, \widetilde{\mathcal{Q}}) = -\log \frac{\exp(\psi(\mathcal{P}_i) \cdot \psi(\mathcal{P}_j))}{\exp(\psi(\mathcal{P}_i) \cdot \psi(\mathcal{P}_j)) + \sum_{k=0}^{n-1} \exp(\psi(\mathcal{P}_i) \cdot \psi(\widetilde{\mathcal{Q}}_k))}, \quad (3)$$

where $\phi(\mathcal{P}) \in \mathbb{R}_+$ is the representation generated by binary classifier, $\psi(\mathcal{P}) \in \mathbb{R}_+$ is the representation generated by multi-class classifier.

Taken together, our final objective function can be written as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{CL}, \quad (4)$$

where λ is a balancing factor for two objectives.

3.3.3 Top-down and Bottom-up Inference

We adopt a top-down clustering and bottom-up pipeline approach for the reordering inference. Specifically, we adopt the Top-down strategy for coarse-to-fine clustering scenes and then shots. And we reorder the different level of videos from bottom to up. Let $X = \langle x_1, x_2, \cdots, x_n \rangle$ be a set of frames in the video clip, with input index set $I = \langle i_1, i_2, \cdots, i_n \rangle$ and GT index set $GT = \langle l_1, l_2, \cdots, l_n \rangle$. We complete the inference in five steps as follows:

Frame cluster to scene We use the output of the scene-level cluster head of HCMC as the feature for k-means clustering. We can obtain the current scene layer clustering groups as $H = \langle h_1, h_2, \cdots, h_m \rangle$, where each group H_k represents the index of the elements in the same scene.

Frame cluster to shot After clustering the scene groups, we use the output of the shot-level cluster head of HCMC as the feature for k-means clustering. At this point, we obtain the grouping of the scene-level clustering as $G = \langle g_{11}, g_{12}, \cdots, g_{mn} \rangle$, where the innermost grouping represents the frames that a shot has.

Frame level reordering After grouping by shot, we can use the frame-level classification head of HCMC to sort the frames in each shot. We can obtain the output vector p between any two frames using a binary classifier, and use the difference in softmax to represent the confidence of the order representation. In this way, we can obtain a weight value between any two frames, and we can obtain an adjacency list. We use beam search to search for the maximum weight arrangement, representing that we have sorted the order of all scenes on a clip using binary classification. At this point, the order becomes $F' = \langle f'_{11}, f'_{12}, \cdots, f'_{mn} \rangle$.

Shot level reordering After reordering frames for each shot, similar to step (3), we sort the shots in each scene. Firstly, we use the concatenated features of each frame in the shot to represent the current shot feature. Then we use the shot-level classification head of HCMC to obtain an adjacency list representing the order confidence. We then use the same beam search method as in step (3) to find the ordering with the maximum total confidence. At this point, the order becomes $G' = \langle g'_{11}, g'_{12}, \cdots, g'_{mn} \rangle$.

Scene level reordering Finally, we sort the scenes by pairing the features of these scenes, using the scene-level classification head of HCMC model to obtain an adjacency list representing the order confidence, and then using the same beam search method to find the ordering with the maximum total confidence. At this point, the scene layer has been ordered as $H' = \langle h'_{11}, h'_{12}, \cdots, h'_{mn} \rangle$. Expanding it into a one-dimensional list represents the order of Pred as $Pred = \langle \hat{l}_1, \hat{l}_2, \cdots, \hat{l}_n \rangle$.

In conclusion, we employed the metric discussed in Sec. 3.2 to compute the accuracy, specifically utilizing the $Pair/TripleScore(Pred, GT)$ formula.

4. Experiments

4.1. Implementation Details

We conduct all benchmark experiments using a single Nvidia 3090Ti. The model is optimized using AdamW [45] with learning rate as $1e-4$ and the following hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-6$. All models are close to convergence with 5 epochs of training. Following [8, 18, 42], our input frames are resized to 224×224. Each image is divided into 32×32 patches before going through the CLIP encoder. To increase the model's robustness, we employ a data augmentation strategy that involves using temporally reversed positive samples. We train the learnable embedding layer from scratch, which features two layers of the BERT transformer block with $hidden\_size = 512$ and $num\_attention\_heads = 8$.

4.2. Main Results on Movie Reordering

We take different types of prevalent pre-trained temporal models as baselines: language model (BERT [44], ALBERT [41]), video model (VIDEOMAE [55]), and video-language model (SINGULARITY [34]). In practice, we follow BERT's next sentence prediction (NSP) task and ALBERT's sentence-order prediction (SOP) task to predict the temporal relation of the input pair by comparing the probability of binary classification. We take the early fusion strategy to fuse the input dialogue and corresponding frames. The overall results are shown in Table 2. The HCMC model with scene or shot clustering achieves state-of-the-art among all the baselines. Whereas, the more fine-grained model, which fuses with both shot and scene information performs much worse than the two-tier joint model. We believe the error accumulation causes this and demonstrate more details in the ablation studies (Sec. 4.4). In addition, the language model performs better than the vision model in general. Figure 7 qualitatively compares the ordering results of ALBERT and our model on randomly selected test data. Compared with the baseline model, our HCMC model demonstrates better shot and scene structure understanding.

	in-domain		out-domain	
	pair.	triplet.	pair.	triplet.
Random	49.66	16.42	49.75	16.59
BERT	53.01	19.44	53.18	19.25
ALBERT	54.06	20.30	53.70	19.76
VIDEOMAE	50.67	18.55	50.83	18.67
SINGULARITY	51.75	18.67	52.33	18.86
HCMC (frame+shot)	**55.40**	**21.58**	**54.97**	**21.23**
HCMC (frame+scene)	54.93	21.43	54.66	20.67
HCMC (frame+shot+scene)	53.10	19.34	52.84	19.13

Table 2. Results on movie reordering.

4.3. Application on Movie Synopsis Association

We test the holistic understanding capability of movie reordering models on Movie Synopses Associations [59] task with the synopsis metadata from MoviePuzzle. Matching the multimodal semantic segments requires the model to have strong long-form video understanding ability. We boost the best-performance video-language pre-trained model SINGULARITY [34] with temporal ordering objective and hierarchical semantic information. Specifically, we continually train the SINGULARITY with multiple tasks, which add the re-ordering task to the original vision-text contrastive task and matching task. Additionally, we fuse the feature of the scene which is extracted from the pre-trained temporal ordering model to the input for the auxiliary. It is worth noting that different from the original temporal ordering model we drop the subtitle for fairness. We evaluate the performance with standard retrieval metrics: recall at rank N (R@N) and median rank (MedR), which measures the median rank of correct items in the retrieved ranking list. As shown in Table 3, we observe that our method boosts the performance of the video-to-text retrieval task more significantly than the

Figure 7. Qualitative comparisons on reordering performance between ALBERT and HCMC (Ours). The red rectangles indicate misplaced single frames, and the yellow ones indicate misplaced shots (the inner frame orders are correct). As seen, our models can better group together frames that share similar shot information to form a plausible visual narrative.

	Movie Segment → Synopsis				Synopsis → Movie Segment			
	R@1	R@5	R@10	MedR(↓)	R@1	R@5	R@10	MedR(↓)
Random	0.13	0.66	1.32	378.5	0.13	0.66	1.32	378.5
zero-shot	0.26	1.06	2.25	230.5	**0.26**	**1.46**	2.65	**299.5**
zero-shot+ours	**0.40**	**1.46**	**3.04**	**220.0**	0.13	1.06	**2.78**	325.5
SINGULARITY	**10.19**	30.16	39.95	18.5	**6.88**	**22.35**	32.94	24.0
SINGULARITY +ours	9.92	**31.88**	**43.39**	**14.0**	5.65	20.63	**33.47**	**22.0**

Table 3. The overall performance of Movie Synopsis Association.

	in-domain		out-domain	
	pair.	triplet.	pair.	triplet.
HCMC (frame+shot)	**55.40**	**21.58**	**54.97**	**21.23**
w/o shot	54.21	20.41	54.29	20.19
w/o CL	52.63	18.79	52.74	18.05
w/o text	50.05	17.53	49.86	17.40
w/o vision	51.69	18.36	51.87	18.67

Table 4. Results on ablated HCMC models.

text-to-video retrieval task. The reason is that our method is helpful in building a long-form video representation while doing no good to the vision-language fusion mechanism. Furthermore, we apply the zero-shot embedding matching method. In practice, we take the clip feature as a baseline and augment it with scene features extracted by our temporal ordering model. Line 2-3 shows zero-shot results similar to SINGULARITY.

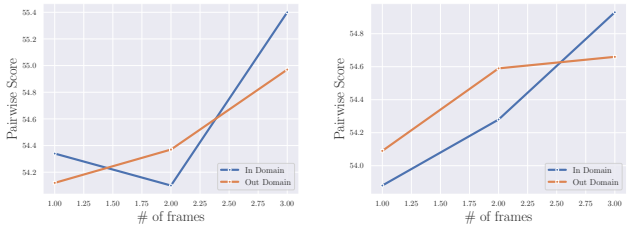
		in-domain			out-domain		
		IoU	pair.	triplet.	IoU	pair.	triplet.
Euclidean	shot	**37.56**	54.93	21.41	**35.74**	54.98	**21.33**
	frame	-	**55.40**	**21.58**	-	54.97	21.23
Cosine	shot	36.9	54.42	20.49	36.70	**55.17**	21.31
	frame	-	54.77	21.15	-	55.10	21.28
Soft_DTW	shot	16.94	52.71	18.04	16.85	53.49	19.83
	frame	-	53.24	19.94	-	54.70	20.94

Table 5. Accumulation error analysis.

4.4. Ablation Studies

Table 4 shows the ablation experiments of different components in the HCMC model. The accuracy of the experiments decreases when any layer shown in the table is removed. Among them, removing the textual information in the clip has the most significant impact on the model, followed by the image information. It can be seen that the MoviePuzzle task relies heavily on multimodal information, which is consistent with our intuition. When contrastive learning or the shot layer is removed from the model, the accuracy of the model also decreases.

Video sampling length We also investigate the impact of the number of hierarchical structures on experimental results through ablation experiments. Figure 8b shows the results of running the frame+shot HCMC model with the lengths of frames within shots controlled at 1, 2, and 3 in the entire video clip. For the in-domain test dataset, the accuracy slightly decreases when the length increases from 1 to 2, but then rapidly increases to its maximum value when the length increases to 3. For the out-domain test set, the model's performance gradually improves with the increasing complexity of the data. The results presented in Figure 8a depict the outcome of utilizing the frame+scene HCMC model to manipulate the lengths of frames within shots at 1, 2, and 3 in the entire video clip. For both the in-domain and out-domain test datasets, the accuracy increases with the increasing length of frames in scenes. From the ablation experiments, it can be concluded that, in general, data with hierarchical structures can be helpful for the model to understand video clips.



(a) Shot Length. (b) Scene Length.

Figure 8. Ablated results on video sampling lengths.

Accumulative Error As mentioned before, our three-tier fine-grained model with both shot and scene clustering modeling obtains poor performance. We believe the accumulative error from the clustering layer and re-order layer causes this. Thus we quantitatively analyze the intermedia result of different stages. Table 5 shows the inter-media score of the shot clustering and frame ordering among different distance functions. We calculate the Intersection-

over-Union (IoU) score of the shot clustering for reference. Specifically, we match the cluster centers with the mean of the ground truth cluster through cosine similarity. The results demonstrate the shot clustering score is positively correlated with the shot ordering score. However, the drop ordering score is more mitigated than the clustering score, *i.e.*, soft_DTW [15]. This is because the soft_DTW method results in more empty clusters than others.

4.5. Human Study

As discussed in Sec. 3.2, the numerical comparison between the predicted order and ground truth may not always be the best metric. To verify our model's inherent understanding of movie logic, we recruited 10 well-educated human testers with proficient English comprehension skills to compare the logical coherence of generated clip sequences. Specifically, we select 20 sets of test data from both in-domain and out-domain sources and fed them into the best-performing baseline model ALBERT and our HCMC model for prediction; see Figure 7 for an exemplar test case. We then present the predicted sequences to testers, who are asked to choose the more reasonable sequence without knowing which model generated it. The orders of the two predicted sequences are randomly shuffled for each comparison test. The comparison results in Table 6 demonstrate that the sequences generated by our approach are more in line with human preferences.

	all	in-domain	out-domain
HCMC (Ours)	**0.55**	**0.54**	**0.56**
ALBERT	0.45	0.46	0.44

Table 6. Human preferences on randomly selected tests.

5. Related Work

Joint representation of images and text Joint image-text representations [6] benefit many language-and-vision tasks by fusion of the modalities. A family of "Visual-BERT" models [53, 12, 3, 41, 44, 65, 21] have been proposed a common method which uses a supervised object detector image encoder backbone and pre-train on image-caption pairs. Cross-modal representations are learned through masked language modeling objective [28] Another family of vision-language models is based on contrastive learning [2, 5, 38, 46, 63], which have a solid ability to extract features on static image-caption pairs. Some other models like Flamingo [1] and Blip-v2 [37] use a lightweight transformer to bridge the modality gap between a frozen image encoder and a frozen large language model (LLM). Our method differs from these approaches as these approaches use an explanate image-text representation which keeps unchanged when learning different semantic structures.

Video-language understanding As an application of artificial intelligence in the multi-media field, video-language understanding has drawn great attention in the research community, such as video story telling [26], video moment retrieval [11], image caption [24, 61], visual question answering [40], and action recognition [58]. Prior arts before the large-scale pre-training era [35, 36, 33] leverage offline extracted video features [27], after that, video-language pre-trained models [39, 68] have shown promising results. Aligned with the success of transformer-based [56] language pre-training models [43, 62], image-text pre-training [41, 23] and video-text pre-training [29, 20] usually use masked visual modeling and have shown promising results on short videos clips. However, through ablation experiments, it was found that these models did not make full use of temporal information on long videos [67, 25].

Movie Benchmarks A couple of research works have addressed the importance of understanding long-form videos, especially movies [49, 22, 30, 7]. For example, MovieQA [54] is extracted from 408 movies with 15K questions. This dataset is designed by using QA to evaluate story understanding. MovieGraph [57] is a small dataset with graph-based annotation of social relationships depicted in clips edited from 51 movies. With the relation graph of character, interaction, and attributions, MovieGraph can offer a hierarchical structure of movie understanding. MovieNet [25] is a holistic movie dataset containing different aspects of annotations which can support comprehensive movie understanding. This dataset covers 1,100 movies, but not all are well annotated and aligned. We carefully re-filter and collate this data set to extract well-labeled and aligned datasets with a movie hierarchy.

6. Discussion and Future Work

This work introduces a novel task **MoviePuzzle**, which aims to facilitate the learning of latent associations between different plot points in movies utilizing recombining multi-modal clips. We curate a new dataset that aligns video clip images and text on a frame-by-frame, sentence-by-sentence basis, accompanied by annotations for the structural hierarchy of frames, shots, and scenes. Furthermore, we devise a Hierarchical Contrastive Movie Clustering (HCMC) model and establish a benchmark for this task.

Limitations Despite the improved performance presented in sufficient experiments, there still remains a huge challenge toward better ordering performance, especially when it comes to long sequences of shuffled frames. One major limitation of the current framework is the lack of global temporal consistency modeling due to the pressure of computational cost. Another limitation lies in the assumption that frames sharing the same shot or scene information shall be grouped together, whereas in a realistic setting, shots may switch back and forth according to the video edit-

ing arts. By presenting MoviePuzzle and these preliminary experiments, we aspire to illuminate future avenues for video comprehension research.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 8

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:25–37, 2020. 8

[3] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2131–2140, 2019. 8

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 1

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *International Conference on Computer Vision (ICCV)*, pages 1728–1738, 2021. 8

[6] Y Bengio, A Courville, and P Vincent. Representation learning: a review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(8):1798–1828, 2013. 8

[7] Subhabrata Bhattacharya, Ramin Mehran, Rahul Sukthankar, and Mubarak Shah. Classification of cinematographic shots using lie algebra and its application to complex event recognition. *IEEE Transactions on Multimedia*, 16(3):686–696, 2014. 8

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 6

[9] Shixing Chen, Xiang Hao, Xiaohan Nie, and Raffay Hamid. Movies2Scenes: Learning scene representations using movie similarities. *arXiv preprint arXiv:2202.10650*, 2022. 3

[10] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2021. 3

[11] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10638–10647, 2020. 8

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Learning universal image-text representations. *CoRR*, 2019. 8

[13] Neil Cohn. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* A&C Black, 2013. 2

[14] Neil Cohn. Visual narrative structure. *Cognitive science*, 37(3):413–452, 2013. 1, 4

[15] Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series. In *International Conference on Machine Learning (ICML)*, 2017. 8

[16] Dave Epstein, Jiajun Wu, Cordelia Schmid, and Chen Sun. Learning temporal dynamics from cycles in narrated video. In *International Conference on Computer Vision (ICCV)*, pages 1480–1489, 2021. 2, 5

[17] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213, 2020. 1

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 6

[19] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET: End-to-end video-language transformers with masked visual-token modeling. In *arXiv:2111.1268*, 2021. 1

[20] Tsu-Jui Fu*, Linjie Li*, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[21] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6616–6628, 2020. 8

[22] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 8

[24] Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. Joint commonsense and relation reasoning for image and video captioning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 10973–10980, 2020. 8

[25] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer, 2020. 1, 2, 3, 8, 12

[26] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1233–1239, 2016. 8

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 8

[28] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019. 8, 12

[29] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 13171–13179, 2021. 8

[30] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008. 8

[31] Hung Le, Nancy Chen, and Steven Hoi. VGNMN: Video-grounded neural module networks for video-grounded dialogue systems. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3377–3393, 2022. 1

[32] Hung Le and Steven CH Hoi. Video-grounded dialogues with pretrained generation language models. *arXiv preprint arXiv:2006.15319*, 2020. 1

[33] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9972–9981, 2020. 8

[34] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 6

[35] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 8

[36] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 1, 8

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 8

[38] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:9694–9705, 2021. 8

[39] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, 2020. 8

[40] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *International Conference on Computer Vision (ICCV)*, pages 10313–10322, 2019. 8

[41] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6, 8

[42] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4804–4814, 2022. 6

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 8

[44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 6, 8

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 5, 8, 12

[47] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10146–10155, 2020. 3

[48] Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–343. IEEE, 2003. 3

[49] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 8

[50] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*, 2019. 2

[51] Dian Shao, Yu Xiong, Yue Zhao, Qingqiu Huang, Yu Qiao, and Dahua Lin. Find and focus: Retrieve and localize video events with natural language queries. In *European Conference on Computer Vision (ECCV)*, pages 200–216, 2018. 3

[52] Vivek Sharma, Makarand Tapaswi, and Rainer Stiefelhagen. Deep multimodal feature encoding for video ordering. *arXiv preprint arXiv:2004.02205*, 2020. 2, 5

[53] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 8

[54] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 8

[55] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 6

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 8

[57] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. MovieGraphs: Towards understanding human-centric situations from videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018. 8

[59] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *International Conference on Computer Vision (ICCV)*, pages 4592–4601, 2019. 6

[60] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057. PMLR, 2015. 1

[61] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10685–10694, 2019. 8

[62] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 8

[63] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 8

[64] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2678–2687, 2016. 1

[65] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 3208–3216, 2021. 8

[66] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[67] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:23634–23651, 2021. 2, 5, 8

[68] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8746–8755, 2020. 8

Supplementary Materials

We provide supplementary materials as follows:

- In Appendix A, we described the steps we took to collect and preprocess our dataset.
- Appendix B presented the experimental hyperparameters and implementation details of our main experiment.
- In section Appendix C, we investigated the impact of clip length on experimental results.
- Appendix D provided a detailed description of our inference algorithm.
- Appendix E introduced alternative model designs and presented experimental findings.
- Appendix F conducted an additional human study experiment.
- In Appendix G we discussed potential ethical risks associated with our paper.
- Appendix H displayed some qualitative result images.

A. Data Collection Information

A.1. Selecting Movie

We employ the following top-level criteria for choosing the movie source and its associated transcripts from MovieNet [25] for **MoviePuzzle**:

- The movies should have color images, and the quality of the picture should not be too dark or blurry.
- The movies should have English subtitles, as some movies in the original MovieNet dataset either have missing subtitles or have subtitles in languages other than English.
- The movies should have synopsis summaries to assist in downstream Movie Synopsis Association tasks. However, it should be noted that only a subset of movies in the original MovieNet dataset have synopsis labels.
- The movies should have a list of actors. We plan to add character labels to the dialogue in future work.
- The movies should have defined boundaries for shots and scenes.

A.2. Aligning Subtitle

We match each frame image with its corresponding subtitle using the following steps:

- Firstly, we remove most voiceovers and background sounds from the subtitles and replace foreign words with English letters with tones.
- Secondly, each dialogue is only paired with images that fall exactly within its corresponding time period, rather than the closest ones.
- Finally, in the case of multiple frames falling on a single dialogue, we merge these frames and select the middle frame as the most representative matching image.

A.3. Cutting into Clips

We segment the movie into clips based on the following criteria, where the textual subtitles are aligned frame by frame:

- The length of each clip is restricted to 10-20 frames.
- The frames containing subtitles in each clip must account for more than 80% of all.
- A greedy algorithm is utilized to match the movie for as long as possible.

After obtaining a total of 10,031 movie clips, we split the data into training, validation, in-domain test, and out-of-domain test sets in the proportions of 70%, 6%, 12%, and 12%, respectively. The dataset will be publicly released at a later date.

B. Additional Implementation Details

The hyperparameters we use are shown in Table 7 below. Our CLIP [46] (clip-vit-base-32[2]) and BERT [28] (bert-base-uncased[3]) parts code uses from HuggingFace.

C. More Numerical Results

We conducted an experiment to compare the impact of different sequence lengths on the performance of our model, as illustrated in the figure below. As the sequence length increases, the overall accuracy of the model demonstrates a declining trend, particularly within the range of 13 to 17 where the decline is more pronounced.

These results suggest that the model's ability to process sequences diminishes as the length of the sequence increases.

D. Inference Algorithm

D.1. Kmeans Clustering Algorithm

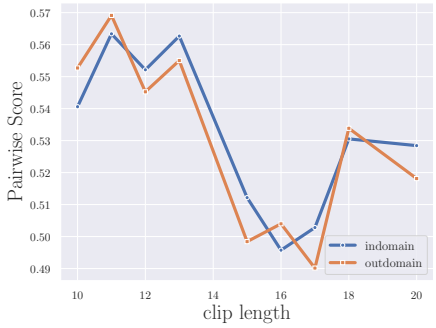
We have a training set $\mathbf{x} = x^{(1)}, \cdots, x^{(n)}$, and want to group the data into a few cohesive clusters. Here, we are

[2] https://huggingface.co/openai/clip-vit-base-patch32

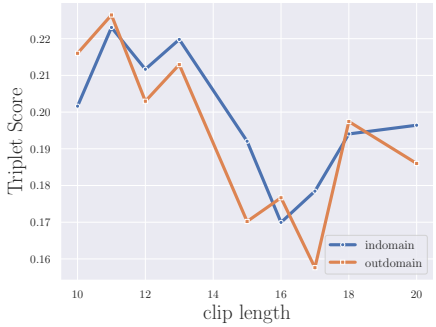
[3] https://huggingface.co/bert-base-uncased

Hyperparam	HCMC
Number of Transformer Layers	2
Hidden Size	512
Attention Heads	8
Attention Heads Size	64
Learning Rate	1e-4
Batch Size	8
Max Cluster Steps	1000
Cluster Distance	Euclidean
Epoch	5
AdamW ϵ	1e-6
AdamW β_1	0.9
AdamW β_2	0.999
Weight Decay	0.01
Patch Size	32

Table 7. Hyperparameters for HCMC.



(a) Pairwise Score.



(b) Triplet Score.

Figure 9. Ablated results on clip length.

given feature vectors for each data point $x^{(i)} \in \mathbb{R}^m$ with no labels as an unsupervised learning problem. Our goal is to predict k centroids and a label $c^{(i)}$ for each datapoint. Then cluster sequence x into k groups based on its corresponding $c^{(i)}$ to obtain a new sequence $y = y^{(1)}, \cdots, y^{(k)}$, where y is a partition of x. The kmeans clustering algorithm is as Algorithm 1:

Algorithm 1 KmeansClustering

Input: training set $\boldsymbol{x} = x^{(1)}, \cdots, x^{(n)}$, cluster number k
Output: label $\boldsymbol{c}$ list

1: Initialize cluster centroids $\mu_1, \mu_2, \cdots, \mu_k \in \mathbb{R}^m$ randomly.
2: For every i, set $c^{(i)} \leftarrow \arg\min_j \|x^{(i)} - \mu_j\|^2$.
3: For each j, set $\mu_j \leftarrow \frac{\sum_{i=1}^n 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^n 1\{c^{(i)}=j\}}$
4: Repeat step 2,3 until convergence or reaching max cluster steps
5: $y \leftarrow$ the partition of x that for all $x^{(i)} \in y^{(j)}$ have same $c^{(i)}$
6: **return** $\boldsymbol{y}, \boldsymbol{c}$

D.2. Beam Search Algorithm

For the training data set $\boldsymbol{x} = x^{(1)}, \cdots, x^{(n)}$, there exists a confidence score between each pair of instances, represented by the adjacency matrix $\boldsymbol{SCORE}$ where $\boldsymbol{SCORE}$ size is $n \times n$ and $\boldsymbol{SCORE}[i, j]$ denotes the weight that x_i is placed directly before x_j in predict sequence. Our goal is to find a path passing through all n points in the complete graph such that the weight of the path is maximized. We utilize beam search to keep track of the top $bsize$ paths at each iteration to search for local optimal solutions. The algorithm is described in Algorithm 2.

Algorithm 2 BeamSearch

Input: $\boldsymbol{SCORE}$(adjacency matrix), $bsize$
Output: $path$

$\boldsymbol{bestList} \leftarrow empty$
$n \leftarrow$ length of $\boldsymbol{SCORE}$
for $begin$ in n **do**
  $\boldsymbol{beamList} \leftarrow \{begin\}$
  $len \leftarrow 1$
  while $len < n$ **do**
    $\boldsymbol{newList} \leftarrow empty$
    for $beamPath$ in $\boldsymbol{beamList}$ **do**
      $\boldsymbol{newList}$.add($\{beamPath + i$, for all node i not in $beamPath\}$)
      $\boldsymbol{newList} \leftarrow$ top $bsize$ score path in $\boldsymbol{newList}$
    end for
    $\boldsymbol{beamList} \leftarrow \boldsymbol{newList}$
    $len \leftarrow len + 1$
  end while
  $\boldsymbol{bestList}$.add(max score path in $\boldsymbol{beamList}$)
end for
$path \leftarrow$ max score path in $\boldsymbol{bestList}$
return $path$

	in-domain		out-domain	
frame	59.75	22.67	59.98	23.06
shot	58.40	21.58	58.97	21.23
scene	58.93	20.43	58.66	20.67
frame+shot	51.21	17.95	51.54	18.01
frame+scene	50.94	18.14	51.49	17.79
frame+shot+scene	50.10	16.54	50.54	16.68

Table 8. Results on MLP training separately.

D.3. Top-down and Bottom-up Inference

For each input test data, after going through the multimodel feature extractor, the input feature representation $X = \langle x_1, x_2, ..x_N \rangle$ is obtained, where N is the length of the video clip and $x_i \in \mathbb{R}^{n \times m}$ where n is the feature length and m is the hidden size. What's more, $\boldsymbol{N_{shot}}$ represents the length of the shot in the scene, and N_{scene} represents the length of the scene. $\phi : \mathbb{R}^{2n \times m} \mapsto \mathbb{R}^2$ is the representation generated by binary classifier, $\psi : \mathbb{R}^{n \times m} \mapsto \mathbb{R}^k$, where k is cluster head's output dimension, is the representation generated by cluster head. We use Algorithm 3 to calculate the score adjacency matrix. Our top-down and bottom-up inference is shown as Algorithm 4

Algorithm 3 OrderScore

Input: X, ϕ
Output: $\boldsymbol{SCORE}$

$n \leftarrow$ length of X
$\boldsymbol{SCORE} \leftarrow n \times n$ empty metric
for i in n **do**
  for j in n **do**
    $\boldsymbol{SCORE}[i, j] \leftarrow \frac{\exp \phi(x_i, x_j)[1] - \exp \phi(x_i, x_j)[0]}{\exp \phi(x_i, x_j)[1] + \exp \phi(x_i, x_j)[0]}$
  end for
end for
return $\boldsymbol{SCORE}$

E. Alternative Design

We have experimented with model structures trained separately at the frame, shot, and scene levels, in addition to the joint end-to-end training structure introduced in the main body of the paper. In essence, we trained three parallel independent models, each having a similar structure. For the Video-Dialogue Encoder layer, we have tried using both a simple MLP structure and a simple transformer block. These single-layer training models achieved satisfactory results in their respective layers (as shown in rows 1-3 of Tables 8 and 9). However, when employing the trained single-layer models for multi-layer inference (as shown in row 4-6 of Tables 8 and 9), the lack of inter-layer connections led to a rapid decrease in model performance.

Algorithm 4 Top-down and Bottom-up Inference

Input: $X, \phi, \psi, N_{scene}, \boldsymbol{N_{shot}}, bsize$
Output: $Pred$

$X \leftarrow KmeansClustering(\psi(X), N_{scene})$
$X0 \leftarrow empty$
for $scene$ in X **do**
  $X0$.add($KmeansClustering(\psi(scene), \boldsymbol{N_{shot}}(i))$)
end for
$X \leftarrow X0$
for $scene$ in X **do**
  for $shot$ in $scene$ **do**
    $\boldsymbol{SCORE} \leftarrow OrderScore(shot, \phi)$
    $path \leftarrow beamSearch(\boldsymbol{SCORE}, bsize)$
    reorder the $shot$ in X according to $path$
  end for
end for
for $scene$ in X **do**
  $\boldsymbol{SCORE} \leftarrow OrderScore(scene, \phi)$
  $path \leftarrow beamSearch(\boldsymbol{SCORE}, bsize)$
  reorder the $scene$ in X according to $path$
end for
$\boldsymbol{SCORE} \leftarrow OrderScore(X, \phi)$
$path \leftarrow beamSearch(\boldsymbol{SCORE}, bsize)$
reorder the X according to $path$
return $Pred$

	in-domain		out-domain	
	pair.	triplet.	pair.	triplet.
frame	61.12	23.01	63.73	23.93
shot	59.43	22.44	59.54	22.25
scene	59.52	21.30	60.12	22.06
frame+shot	51.99	18.12	52.28	18.67
frame+scene	51.54	17.55	51.83	18.32
frame+shot+scene	50.65	17.06	51.52	17.11

Table 9. Results on Transformer training separately.

F. Human Study

To verify our model's inherent understanding of movie logic, we further recruited 10 well-educated human testers with proficient English comprehension skills to compare the logical coherence of generated clip sequences with ground truth. Similar to the main body of the paper, we select 20 sets of test data from both in-domain and out-domain sources and ask the testers to choose the more reasonable sequence without knowing whether it predicted by our HCMC model. The comparison results in Table 10 demonstrate that the sequences generated by our model still have a probability of about 26.5% of being selected as more reasonable, indicating that the sequences generated by our model may have the same potential Visual Narrative Structures as the ground truth.

	all	in-domain	out-domain
HCMC (Ours)	26.4%	22.8%	30%
Ground Truth	73.6%	77.2%	70%

Table 10. Human preferences on randomly selected tests.

G. Ethics Concern

Were any ethical review processes conducted (e.g., by an institutional review board)? No official processes were done, as our research is not on human subjects, but we had significant internal deliberation when choosing the movies.

Does the dataset contain data that might be considered confidential? No, our data comes from existing public movie data sets.

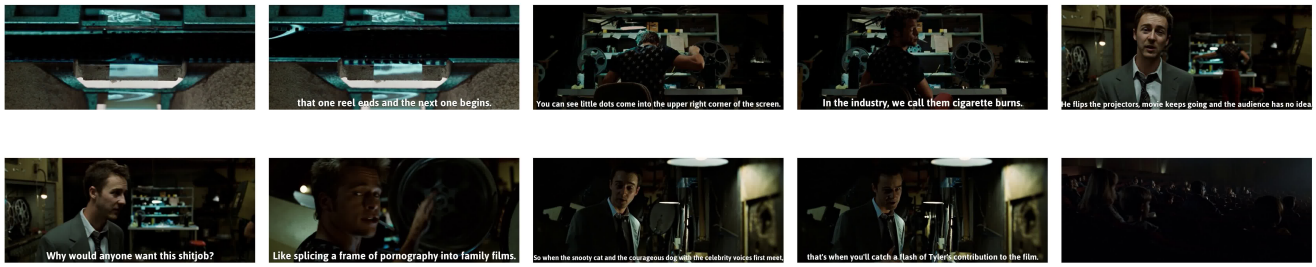
Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. Yet – few of these videos exist; There may be horror films and things like profanity, but all the films have been censored for public screening.

Does the dataset identify subpopulations (e.g., by age or gender)? Not explicitly.

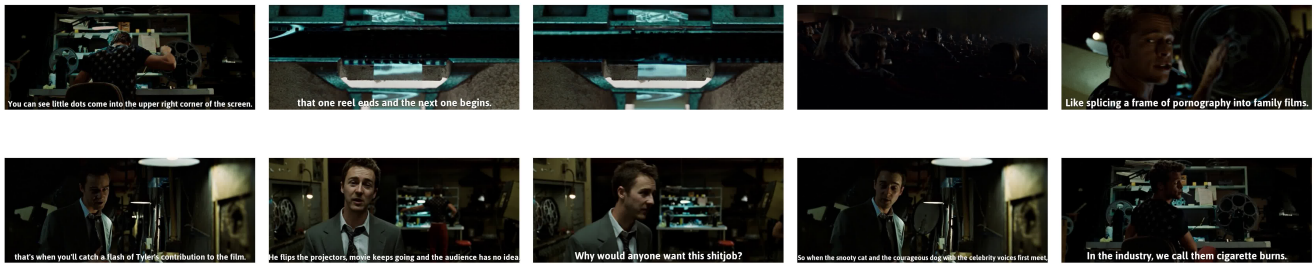
Is it possible to identify individuals (i.e., one or more natural persons) directly or indirectly (i.e., in combination with other data) from the dataset? Yes, our data include celebrities or other film actors. All of the videos we use are publicly available datasets.

H. Qualitative Image

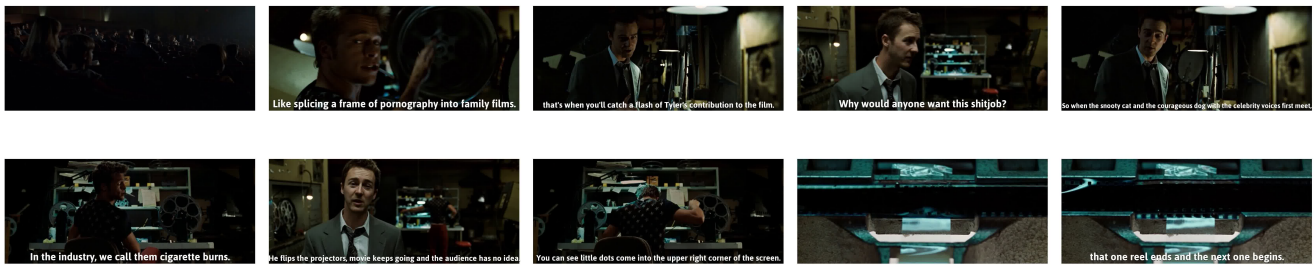
Refer to Figures 10 to 15 for qualitative comparisons. Figures 10 to 12 come from in domain test data set and Figures 13 to 15 come from out domain test dataset. From these examples, it can be observed that our method is able to arrange the same shots closer together within a movie clip compared to the baseline, resulting in an overall better reordering performance.



(a) Ground Truth

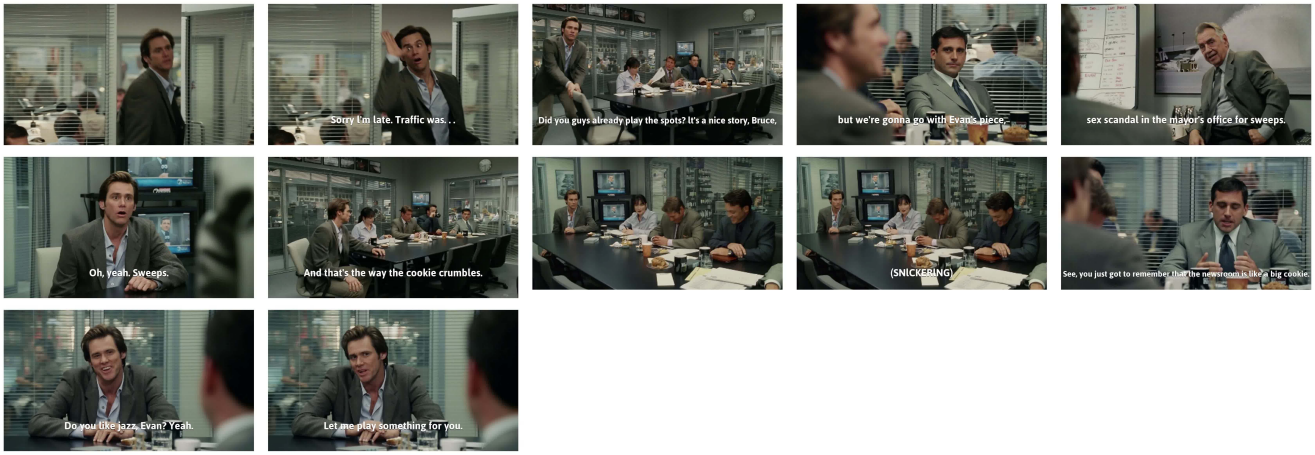


(b) HCMC

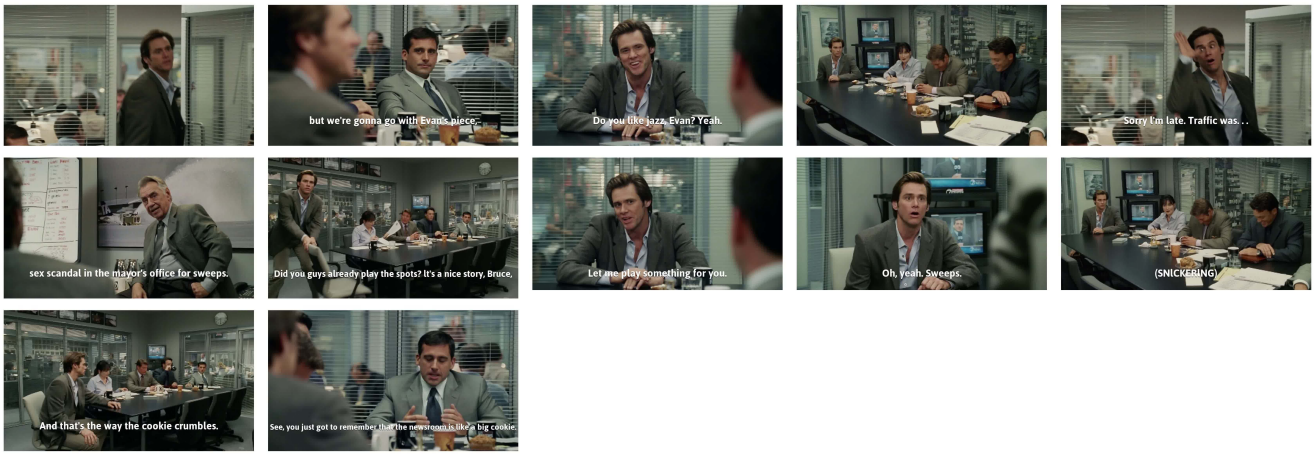


(c) Baseline

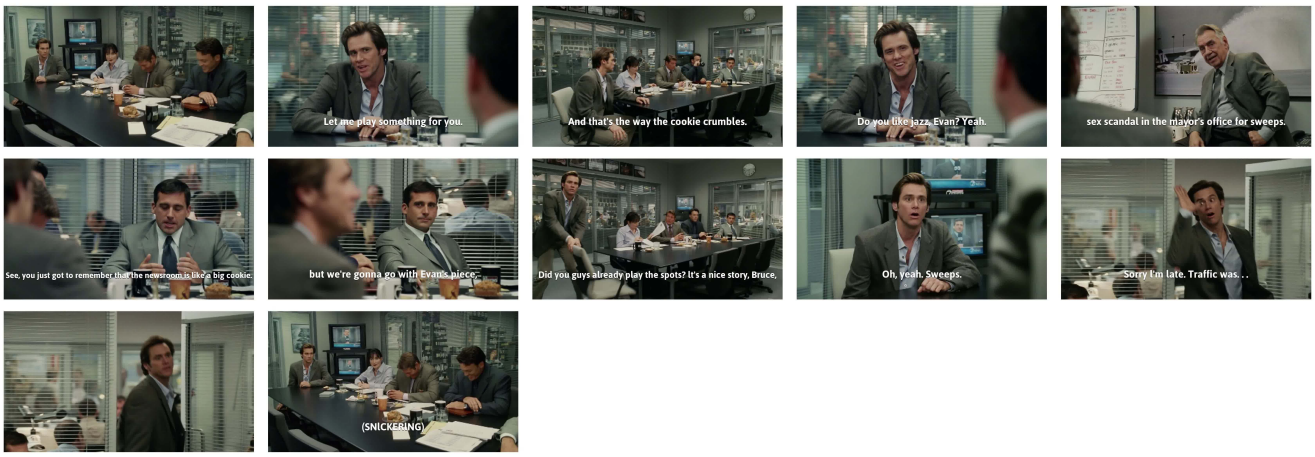
Figure 10. **Movie:** *Fight Club*. **Synopsis:** After one meeting, he confronts her. She argues that she's doing exactly what he does and quips that the groups are 'cheaper than a movie and there's free coffee. Instead of ratting each other out, they agree to split up the week and exchange numbers. Despite his efforts, the narrator's insomnia continues. On a flight back from one of his business trips, the narrator meets Tyler Durden. Tyler offers a unique perspective on emergency procedure manuals in the plane and they strike up a casual conversation. Tyler is a soap salesman, if he's not working nights as a projectionist and slipping bits of porn between reels. The narrator arrives at the baggage claim to discover that his suitcase has been confiscated, most likely due to a mysterious vibration, before he taxis home. However, home, a fifteenth story condominium, has been blasted into the night by what was theorized to be a faulty gas line ignited by a spark on the refrigerator. Having nowhere to go, the narrator finds a business card for Tyler and calls him up. They meet in a parking lot behind a bar where Tyler invites the narrator to ask to come live with him...on one condition: that the narrator hit Tyler as hard as he can. The narrator, though puzzled, complies and they engage in a fist fight before sharing a couple of drinks. The experience is surprisingly euphoric.



(a) Ground Truth

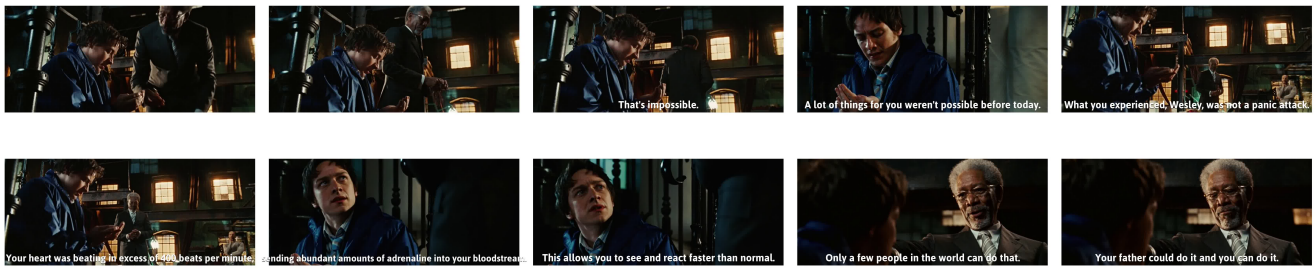


(b) HCMC

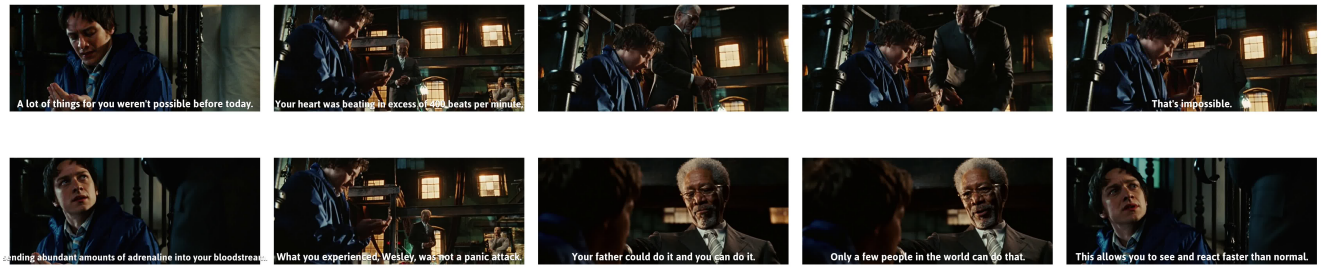


(c) Baseline

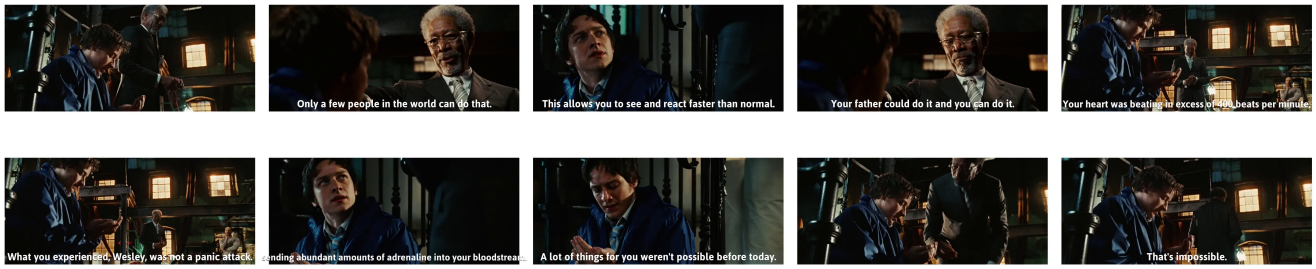
Figure 11. **Movie:** *Bruce Almighty*. **Synopsis:** Racing back to the station, Bruce gets caught in traffic and vents his frustration about the fact that his life is in a go-nowhere rut. Arriving late to an important meeting, fellow staffers–including nemesis Evan Baxter (Steve Carrell)–needle Bruce mercilessly about his clownish coverage at the bakery, further exacerbating his bitterness about being stalled on his career path. In the exchange with Evan, we see that Bruce has a lively, but dark, sense of humor and won't take anything lying down. After the meeting, Bruce begs his boss, Jack Baylor (Philip Baker Hall), to consider him for the open anchor position.



(a) Ground Truth



(b) HCMC

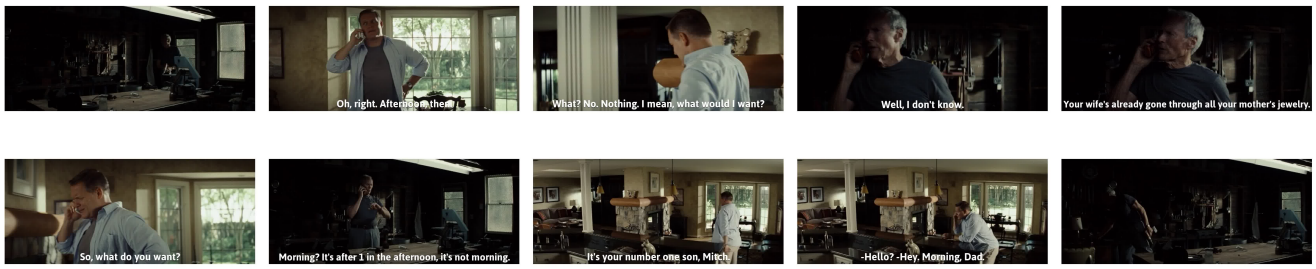


(c) Baseline

Figure 12. **Movie:** *Wanted*. **Synopsis:** One night at a pharmacy, Gibson meets a mysterious woman who tells him his father was an elite assassin who had been killed the day before. Gibson replies that his father abandoned him a week after his birth. At that moment, Cross appears, gun in hand. The woman opens fire on Cross. Gibson and the woman escape from the resulting shoot-out and have a wild car chase in the streets of Chicago. The woman brings Gibson to the headquarters of The Fraternity, a thousand-year-old secret society of assassins. The group's leader, Sloan (Morgan Freeman), formally introduces Gibson to Fox (Angelina Jolie), the woman from the night before, and invites him to follow in his father's footsteps as an assassin. Sloan tests Gibson by making him shoot the wings off a fly. When Gibson refuses, a gun is put to his head, triggering a panic attack. Gibson somehow manages to shoot the wings off several flies. Sloan says that he was able to do that because his heart beats 400 times a second when he's stressed. When Sloan asks him whether he want to know how to control it, he runs away in fear. Gibson wakes up the next day hoping everything was a dream, but discovers his father's gun (which he stashes in the toilet tank), and that he has $3.6 million in his bank account. At work, Gibson tells off his boss, bashes his duplicitous friend with a computer keyboard, and storms out. Gibson then sees pictures of himself and Fox on the front page of several newspapers as wanted fugitives for the pharmacy shooting. Then he notices Fox, who has been waiting outside, and she gives him a ride back to the Fraternity headquarters - an unassuming textile mill.

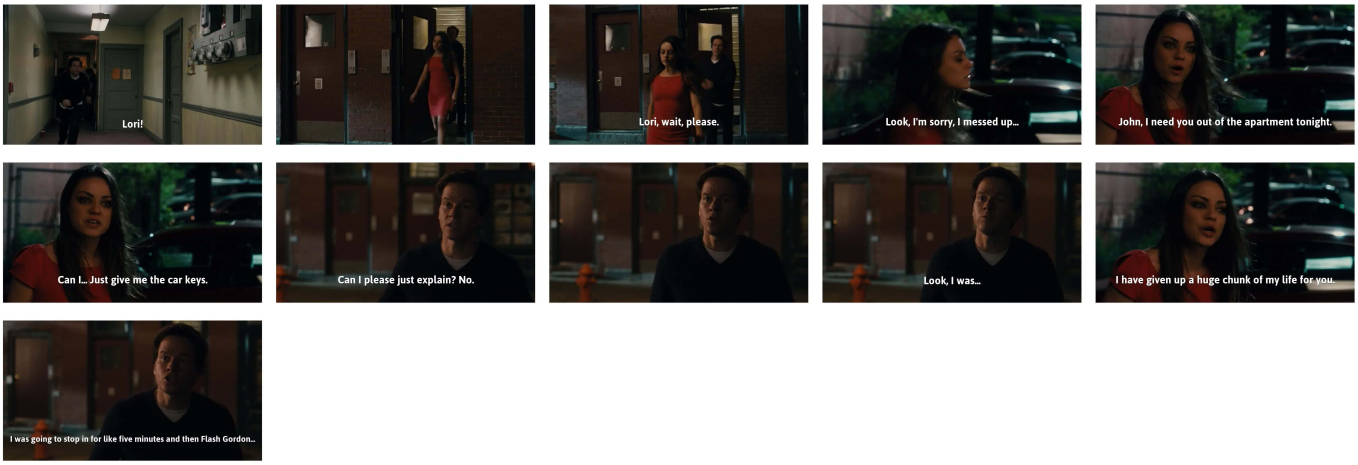
(a) Ground Truth

(b) HCMC

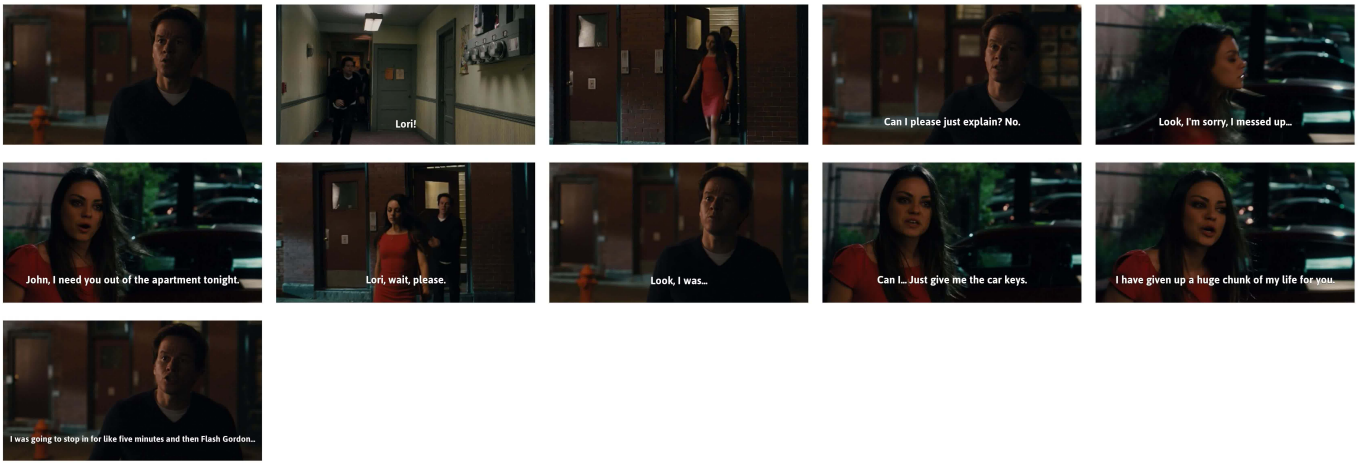


(c) Baseline

Figure 13. **Movie:** *Gran Torino*. **Synopsis:** Mitch and his wife, Karen (Geraldine Hughes) go to visit Walt on his birthday, bringing him a cake and a few gifts meant to make certain menial tasks easier. Presentation, and explanation, of these gifts quickly turn into a shamelessly brazen pitch to get Walt to move into a senior's retirement home. Knowing that Mitch and Karen just want to get their hands on his house, Walt growls in anger and throws them out; gifts, cake and all. Mitch and Karen cannot understand Walt's reaction.



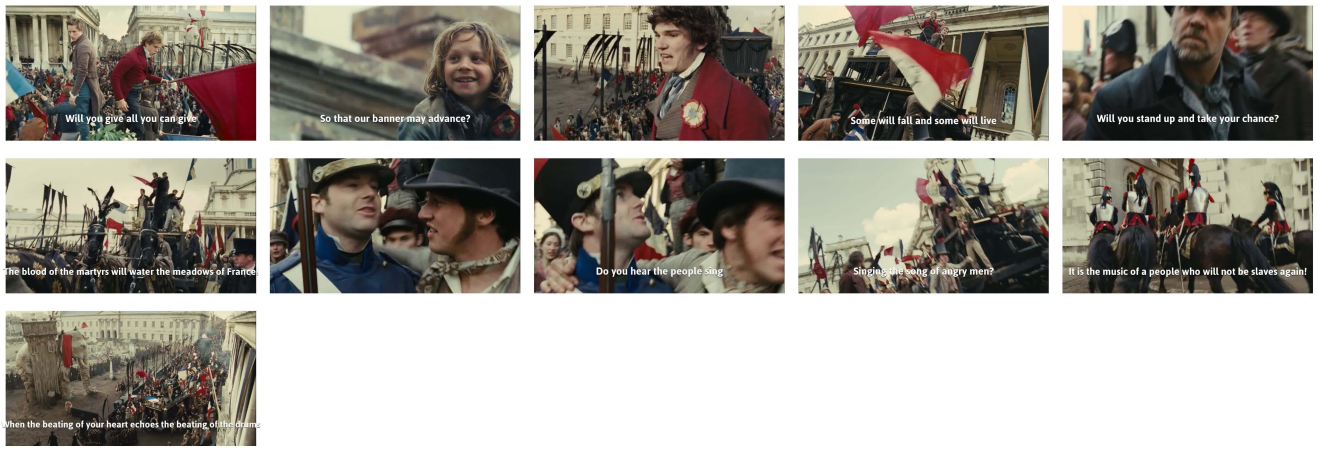
(a) Ground Truth



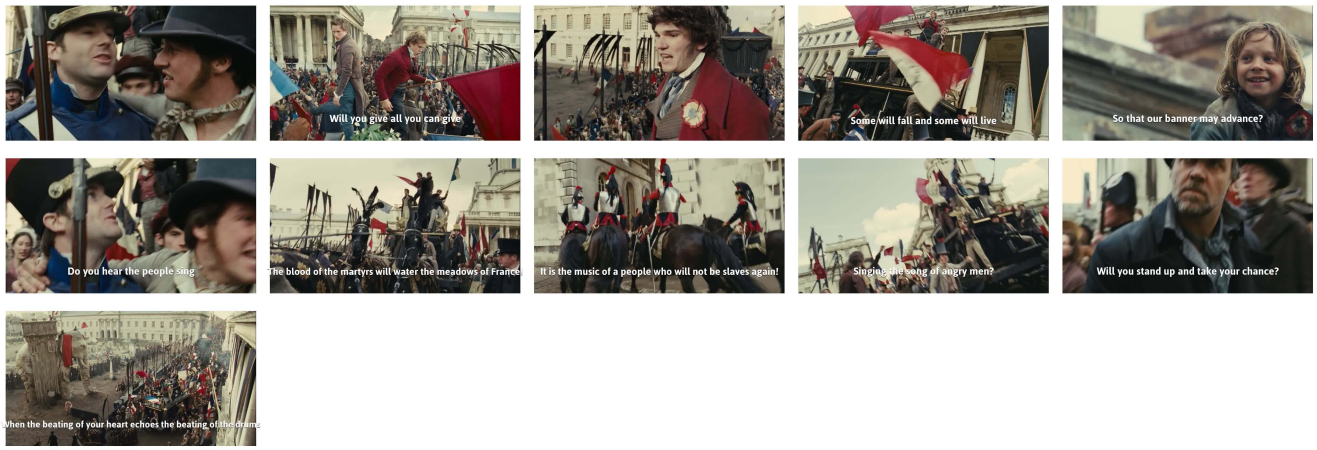
(b) HCMC

(c) Baseline

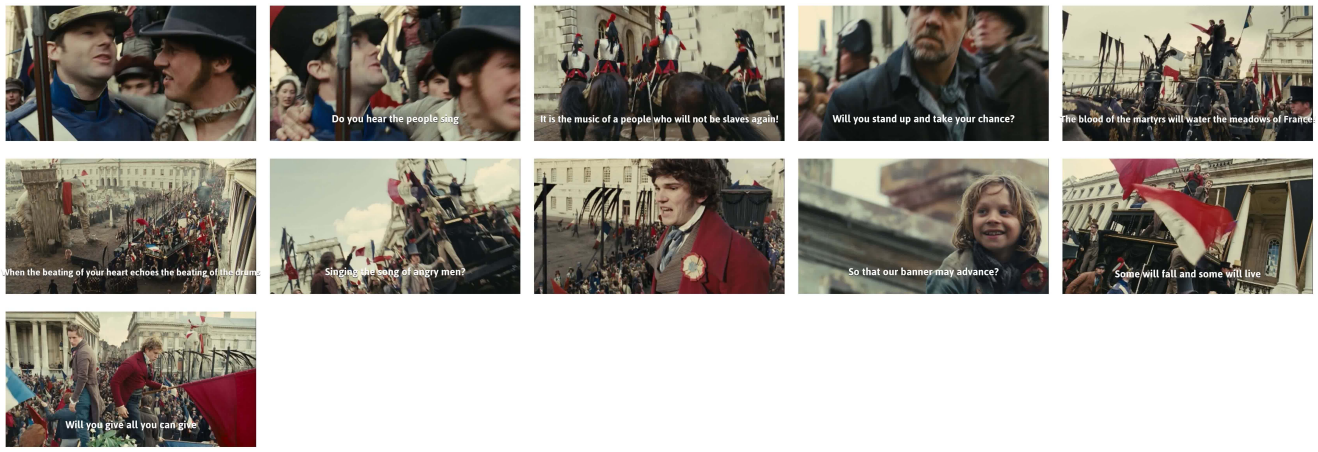
Figure 14. **Movie:** *Ted*. **Synopsis:** John finds Ted his own apartment and a job at a grocery store, where his grossly irresponsible behavior on the job manages to both get him promoted and acquainted with the superficial co-worker Tami-Lynn (Jessica Barth), who gets easily irritated by Lori who is shocked at her anger. Regardless, Ted and John still spend most of their time together, which frustrates Lori when she discovers John has been skipping work to do so while using her for his excuses. Meanwhile, a crazed loner named Donny (Giovanni Ribisi), who idolized Ted as a child, shows interest in possessing him for his brutishly destructive son, Robert (Aedin Mincks). Things start to come to a head when Lori and John are invited to a party put on by Lori's lecherous manager, Rex (Joel McHale), and Ted lures John away to a wild party at his apartment with the offer to meet Sam J. Jones (playing himself), the star of their favorite movie, Flash Gordon. Although John arrives with the intention of spending only a few minutes, he gets caught up in the occasion which gets completely out of control, with Sam J. Jones persuading John and Ted to snort cocaine and Ted singing karaoke and eventually getting beaten-up by a duck. Eventually, Lori discovers John there and breaks up with him in a rage. At that, John blames Ted for ruining his life and tells him to stay away.



(a) Ground Truth



(b) HCMC



(c) Baseline

Figure 15. **Movie:** *Les Misérable*. **Synopsis:** The next day, the students interrupt Lamarque's funeral procession and begin their assault. Javert poses as a rebel in order to spy on them, but is quickly exposed by Gavroche and captured. During the ensuing gunfight, Eponine saves Marius at the cost of her own life, professing her love to him before she dies, which leaves Marius devastated at the loss of his best friend. Valjean, intercepting a letter from Marius to Cosette, goes to the barricade to protect Marius. After saving Enjolras from snipers, he is allowed to execute Javert. However, when the two are alone, Valjean chooses to free Javert instead and fires his gun to fake the execution. Initially disbelieving, Javert wonders at Valjean's generosity.