MMAD: Multi-modal Movie Audio Description

Xiaojun Ye[1], Junhao Chen[2], Xiang Li[3], Haidong Xin[4], Chao Li[5], Sheng Zhou[1*], Jiajun Bu[1]
[1]Zhejiang University, [2]Tsinghua University, [3]Peking University
[4]Northeastern University, [5]Harbin Engineering University
{xiaojunye81, yisuanwang, lx1906815508, hdxin2002}@gmail.com
lichao006@hrbeu.edu.cn
{zhousheng_zju, bjj}@zju.edu.cn

Abstract

Audio Description (AD) aims to generate narrations of information that is not accessible through unimodal hearing in movies to aid the visually impaired in following film narratives. Current solutions rely heavily on manual work, resulting in high costs and limited scalability. While automatic methods have been introduced, they often yield descriptions that are sparse and omit key details. ddressing these challenges, we propose a novel automated pipeline, the Multi-modal Movie Audio Description (MMAD). MMAD harnesses the capabilities of three key modules as well as the power of Llama2 to augment the depth and breadth of the generated descriptions. Specifically, first, we propose the ***Audio-aware Feature Enhancing Module*** to provide the model with multi-modal perception capabilities, enriching the background descriptions with a more comprehensive understanding of the environmental features. Second, we propose the ***Actor-tracking-aware Story Linking Module*** to aid in the generation of contextual and character-centric descriptions, thereby enhancing the richness of character depictions. Third, we incorporate the ***Subtitled Movie Clip Contextual Alignment Module***, supplying semantic information about various time periods throughout the movie, which facilitates the consideration of the full movie narrative context when describing silent segments, thereby enhancing the richness of the descriptions. Experiments on widely used datasets convincingly demonstrates that MMAD significantly surpasses existing strong baselines in performance, establishing a new state-of-the-art in the field. Our code will be released at https://github.com/Daria8976/MMAD.

Keywords: Multi-modal Learning, Audio Description, Caption Generation

1. Introduction

Cultural productions are increasingly integrating the visually impaired due to evolving legal requirements and the growth of societal support(Han et al., 2023). It's a well-known fact that movie stands as a prevalent art form. Yet, the absence of voice narration means many aren't tailored to the disabled. Films made accessible for disabled viewers have been adapted for their benefit. For the visually impaired, films require voice-over narrations to describe non-dialogue scenes(Wikipedia contributors, 2023). This voice-over process, called Audio Description (AD), describes the movie's visual components(Han et al., 2023) for BVI.

The conventional process of creating accessible movies relies heavily on manual work, leading to high costs and lengthy production times, making scalability a challenge(Han et al., 2023). The automated generation of AD constitutes a multi-modal translation task. Specifically, this technique relies on computer vision techniques for video content analysis and segmentation to recognize visual information, such as important objects(Robinson et al., 2020), relationships between objects(Kukleva et al., 2020), and their actions and interactions(Patron-Perez et al., 2010; Vondrick et al., 2016). The audio features obtained by the audio encoder(Guzhov et al., 2022) are then concatenated as input for the natural language processing module. Natural language processing techniques can generate vivid descriptions of video content based on visual and auditory information, using vocabulary that adheres to linguistic expressions(Li et al., 2022).

Despite its importance, the vision community hasn't extensively studied AD. Automatic AD creation differs from typical vision-to-language tasks, bringing forth unique challenges. Crucially, AD for a given video clip considers several factors: visual cues, prior ADs and subtitles (linguistic context), audio cues, and time. The model's expected outcome is a cohesive cross-modal embedding(Koepke et al., 2023). Furthermore, ADs omit descriptions of scenes understandable from background noises and are strategically timed not to coincide with dialogue. Unlike generalized descriptions provided by dense video captioning(Iashin and Rahtu, 2020), AD offers specific details, identifying characters and their actions.

Our primary contributions include:

- Based on the complementary nature of multimodal semantic information, we propose a novel framework that is adept at utilizing multiple modal inputs to enhance AD generation.

*Corresponding author

LREC-COLING 2024, pages 11415–11428

Sole reliance on audio content can usher in semantic discrepancies due to its inherent ambiguity (Drossos et al., 2020). Diverging from prevalent AD generation frameworks, we've incorporated an ambient music input modality. This addition aims to offer visually impaired individuals a richer information spectrum.

- We design the narrator interval detection module for pinpointing proper time intervals for AD insertion incorporated both speech and text recognition into.
- For multimodal inputs, we design the Audio-aware Feature Enhancing Module, the Actor-tracking-aware Story Linking Module, and the Movie Clip Contextual Alignment Module using a unimodal training method, and design a multimodal converter in the input layer of the framework to realize multimodal fusion.
- We've scrutinized MMAD's capabilities both quantitatively and qualitatively across demanding datasets. Additionally, we design human evaluation to evaluate its performance on real-world movie clips. The experimental results demonstrate that MMAD exhibits a superior level of both information utilization and generalization compared to existing techniques.

2. Related Work

Multimodal video subtitles. The ADLAB PRO guide's survey results on the needs of visually impaired groups in various countries for AD point out that AD requires accurately convey the plot and details of movies or other cultural events. Drawing inspiration from dense image captioning paradigms(Johnson et al., 2016), Krishna and team pioneered the dense video captioning arena, underpinned by the ActivityNet Captions dataset(Krishna et al., 2017; Zhou et al., 2018; Mun et al., 2019; Rahman et al., 2019). Vladimir & Esa's MDVC approach(Iashin and Rahtu, 2020) proffered an amalgamation of modalities, underscoring that audio-visual synergy enhances video caption quality. Video subtitles are responsible for obtaining important elements, relative relationships and action behaviors in video key frames, which constitute the main part of the movie audio description.

In addition to the video module, a good AD needs to provide as much information as possible for description optimization. The inaugural venture into audio captioning surfaced in (Drossos et al., 2017), utilizing PSE's auditory datasets and harnessing BiGRU(Rana, 2016)-centric models. Subsequently , endeavors like those by (Xu et al., 2021) dissected audio subtitle semantics within a comprehensive framework, marking SOTA milestones by adeptly weaving in diverse informational threads via transfer learning. Apart from improvements in the modal input part, the emergence of large language models has brought huge improvements to AD quality compared to previous models. Inspired by Vision-LLM(Wang et al., 2023) and AnyMAL(Moon et al., 2023), the MMAD proposed in this paper maps multi-modal features into a language-aligned feature space, and uses LLama(Touvron et al.) decoding to obtain the final AD.

Video subtitles for BVI. Unlike traditional video subtitles, accessibility-oriented video subtitles need to meet the specific needs of visually impaired viewers. This type of accessible audiovisual media working model is expected to comply with accessibility regulations and meet the needs of the visually impaired community for audio description. Furthermore, compared with the independent and separated video caption model, movie audio description needs to maintain the memory of the previous content for maintaining the smoothness of the narrative.

Wang et al.(Wang et al., 2021b) proposed an end-to-end system for automatic audio description generation.The system utilizes an attention-based video dense caption generation model to generate descriptions for all events in each inconsistent video clip. However, the challenge of creating contextually rich and timely descriptions remained. Vander Wilt and Farbood(Vander Wilt and Farbood, 2021) tackled live theater accessibility, proposing an online time warping algorithm for aligning pre-recorded audio descriptions. Their approach was innovative but faced challenges in handling the dynamic nature of live performances. To further this work, Rocha Filho et al.(Filho et al., 2021) introduced a system for automatic character description in videos, employing deep learning techniques. However, the challenge of seamlessly integrating these descriptions into the video narrative persisted. Finally, Campos et al.(Campos et al., 2023) explored CineAD, a system for audio description using movie scripts and visual information. Despite its potential, it struggled with synchronizing descriptions with live video content.

Previous researchers mainly focused on accessible video subtitles and lacked contextual information integration modules suitable for movie-level audio description. Therefore, we designed the Movie Clip Contextual Alignment Module to provide more contextual information in movie audio description. In addition, No previous work has applied multimodal techniques to movie descriptions, so previous audio descriptions generated by automated systems tended to be single descriptions that lacked emotional coloring. This article proposes the Audio-aware Feature Enhancing Module, which uses LLM to generate richer and smoother audio descriptions in a multi-modal manner. Most importantly, previous work lacked a suitable character recognition module. The Actor-tracking-aware Story Linking Module proposed in this paper is suitable for character recognition in movie scenes where faces are often missing, it helps the visually impaired group better understand the plot of the movie.

3. Method

3.1. Overview

Indeed, Audio Description (AD) plays a paramount role in enhancing film accessibility, providing crucial information that can't be adequately conveyed through dialogue alone. Given a comprehensive movie denoted as V, we decompose it into several shorter segments, represented by $x_1, x_2, \ldots, x_N$. The initial step involves pinpointing proper time intervals for AD insertion. During these segments, generate AD with proper length that does not interfere with the original audio in the films. To help BVI understand the story flow and get comprehensive information, the generated AD contains characters' name and meaningful background music.

To address the above needs and challenges, we propose MMAD pipeline, as illustrated in fig. 1, we leverage three key modules to generate effective movie audio description. First, we employ the Audio-aware Feature Enhancing Module. This module doesn't focus on spoken words, but rather, it centers on background sounds and music to deliver atmosphere, mood, and environmental information. Second, we have the Actor-tracking-aware Story Linking Module. This module accurately links the active character from multiple angles by replacing personal pronouns in the caption with specific character names to provide a clearer narrative context. Lastly, we use the Movie Clip Contextual Alignment Module, which supplements dialogue scenes with information from other audible segments to ensure comprehensive description generation. The outputs from these three modules, combined with the multimodal inputs, are mapped to the textual embedding domain of a specific Large Language Model LLaMA2 (Touvron et al., 2023a). By merging the word embeddings from earlier movie audio descriptions and subtitles, these serve as prompts for the expansive language model to generate the final extensive description. We will now delve into the detail of each module.

3.2. Movie Clip Contextual Alignment

We categorize movies into dialogue segments and non-dialogue segments. As previously mentioned, generated AD can't overlap with the dialogue segments. Thus, the current movie audio description methods typically focus solely on the visual information present in non-dialogue segments, which, however, ignores the visual details not provided in the main video audio track during dialog segments. To address this challenge and ensure a coherent, context-aware description, we propose the Movie Clip Contextual Alignment Module in MMAD.

The first step of the module is to identify appropriate intervals for inserting movie audio description. In order to detect non-dialogue segments, we employ WhisperX(Bain et al., 2023) to extract character dialogues and convert them into textual subtitles. Subsequently, the Connectionist Temporal Proposal Network (CTPN)(Tian et al., 2016) is utilized to identify sequences of frames without subtitles, marking these as potential intervals for movie audio description insertion.

Upon identifying the intervals for movie audio description insertion, the module generates contextually relevant descriptions for both dialogue and non-dialogue frames within these intervals. This process considers not just the visual information from the current clip, but also the content from previously captioned clips.

The visual mapping network $\mathcal{M}_V$, which takes as scene-specific frame features from the current movie clip x_i as prefix inputs to the language model:

$$h_{x_i} = \mathcal{M}_V\left(\{z_1, \cdots, z_N\}\right); z_i = f_{\text{CLIP}}\left(x_i\right) \quad (1)$$

The length limit of the generate AD as follows:

$$\left(h_{x_i^{(B)}} + h_{x_{i+1}^{(C)}}\right)/r \leq t\left(x_{i+1}^{(C)}\right), x_i^{(B)} \notin x_{i+1}^{(C)} \quad (2)$$

Here, r represents frame rate, and $t(x_{i+1}^{(C)})$ is the time duration of the non-dialogue clip. This limit is used to iteratively optimize the generated AD during training, which ensures the most significant visual information is relayed within the given time constraints and maintains the narrative flow and context of the movie, significantly enhancing the description generation quality.

3.3. Actor-tracking-aware Story Linking

Unlike normal video captioning tasks, in order to help visually impaired people understand the storyline, the generated narration should be referred to by the name of a specific character, not by pronouns such as "he", "she", "it" or other pronouns. How to match the active character in the current movie clips is a challenging task. In order to achieve this function and improve the quality of the generated AD, we(i)input an actor identity matching table, which contains the actor's character photo and corresponding character name in the film or TV, in addition to the film clips, and input the matching table to our designed character portrait feature

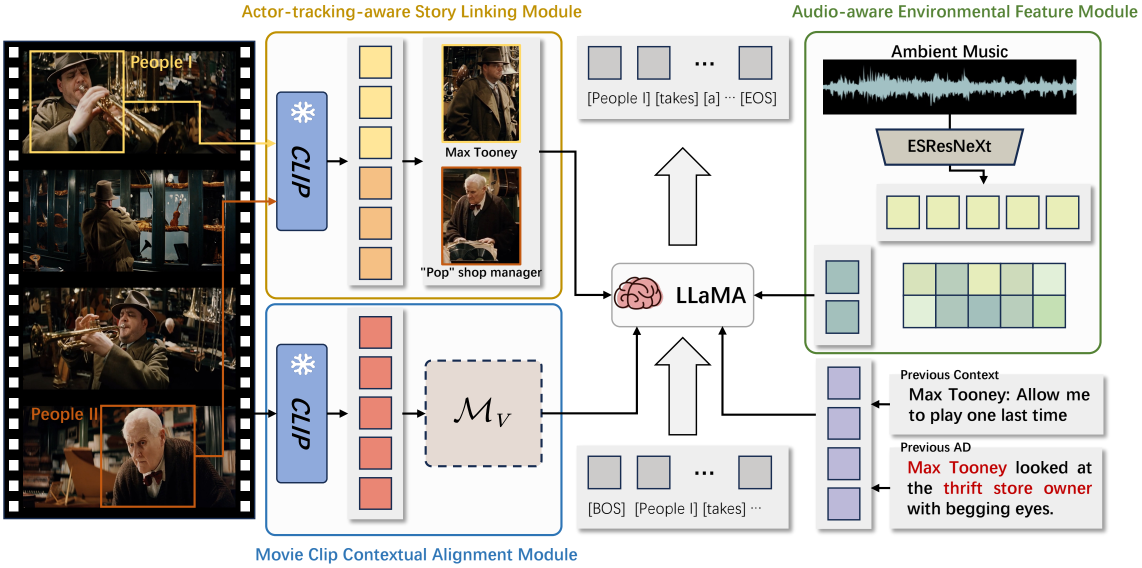

Figure 1: Overview of MMAD: MMAD consists of multiple modality encoders used to generate movie narration

calibration module for learning visual character features; (ii)establish a character recognition module that leverages these learned visual character features. This module matches characters within current movie clips to their corresponding visual features. The design of the module is shown in fig. 2.

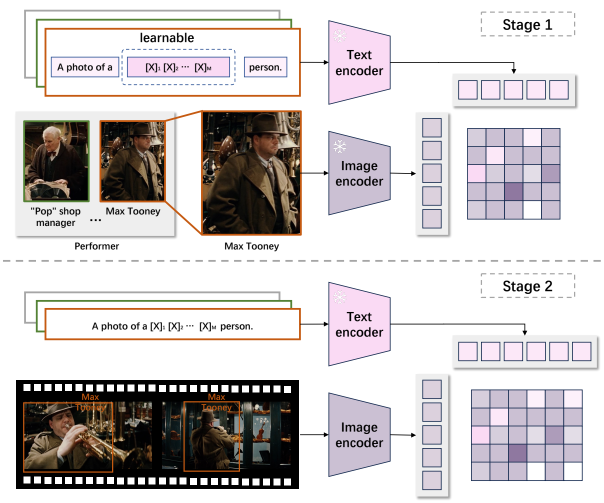

Figure 2: Overview of our Actor-tracking-aware Story Linking Module, which fixes the text encoder and image encoder in the first training stage, optimizes a set of learnable text tokens to generate the text features, and then uses the text features to optimize the image encoder in the second stage.

3.3.1. Character Portrait Features Calibration

Before we proceed to the character recognition stage, it is essential to calibrate the character portrait features. This step is necessary because cinematic depictions present a range of complexities, such as changing viewing angles and occasional partial obscurity of characters. To address these challenges and ensure precise recognition, we integrate Image Re-identification (ReID) techniques (Li et al., 2023a) into our model.

To achieve this, our model adopts a contrastive learning approach. We introduce ID-specific learnable tokens that help interpret ambiguous textual descriptions and calculate a contrastive loss between character images and texts.

Specifically, we define the character image-to-character text contrastive loss ($\mathcal{L}_{i2t}$) as:

$$\mathcal{L}i2t = -\sum i = 1^N log \frac{exp(\mathbf{f}_i \cdot \mathbf{g}i/\tau)}{\sum j = 1^N exp(\mathbf{f}_i \cdot \mathbf{g}_j/\tau)} \tag{3}$$

where, $\mathbf{f}_i$ and $\mathbf{g}_i$ represent the i-th image and text feature vectors, respectively. τ represents a scalar temperature, while N denotes the batch size (Deng et al., 2020a). Through this design, our model is better equipped to handle the complex visual narratives presented in films.

3.3.2. Character Recognition

Once we've calibrated the character features, we transition to the character recognition stage. In this phase, we optimize the parameters in the image encoder using a combination of triplet loss and ID loss with label smoothing for optimization (He et al., 2021). This optimization process enhances the model's ability to accurately recognize and differentiate between various characters.

In addition to these loss functions, we also design a cross-entropy loss from image features to text features, denoted as $\mathcal{L}_{i2tce}$. This loss is defined as:

$$\mathcal{L}_{i2tce} = -\sum_{i=1}^{N} \mathbf{p}_i log(\mathbf{q}_i) \tag{4}$$

In this equation, $\mathbf{p}_i$ represents the ground truth distribution, and $\mathbf{q}_i$ denotes the predicted distribution (Chollet, 2017).

3.4. Audio-aware Feature Enhancing

In movies, environmental sounds are a rich source of information and often greatly influence our interpretation of a scene. For instance, the same visual scene could convey entirely different meanings with a background score of eerie suspense compared to a cheerful melody. In addition, audio information is ambiguous, and the inclusion of audio core information in the narration is necessary to aid BVI people's understanding of the story of the movie. To address this need, we designed the Audio-aware Feature Enhancing Module. This module specifically targets the extraction and enhancement of salient audio features from environmental sounds in a movie, providing additional cues for generating precise visual descriptions. It comprises an Ambient Audio Encoder(section 3.4.1) and a Modality Alignment Module(section 3.4.2).

3.4.1. Ambient Audio Encoder

To effectively utilize the audio cues in movies, we first need to obtain powerful auditory features that can effectively encapsulate the rich information embedded in the audio track. The audio features not only need to capture the raw attributes of the audio signal but also need to highlight the semantic and contextual aspects that can supplement the visual cues in the movie scenes.

In our quest to capture these potent audio features, we adopt ESResNeXt (Donahue et al., 2015) as our audio encoder. ESResNeXt is constructed on the efficient ResNeXt (Chollet, 2017) backbone network and includes a trainable time-frequency transformed fbsp layer. This unique layer, inspired by complex-frequency B-spline wavelets (Teolis and Benedetto, 1998), optimizes the time-frequency representation of sound through end-to-end learning. More specifically, we employs the Short-Time Fourier Transform (STFT) to transmute raw audio signals into time-frequency representations as per the following equation:

$$X(x,\tau) = \sum_{n=-\infty}^{\infty} x[n]w[n-\tau]K_{f_c}^{DFT}(n) \tag{5}$$

In this way, we manage to capture strong and meaningful audio features that enhance the MMAD model's ability to generate precise and contextually accurate movie descriptions.

3.4.2. Modality Alignment Module

Once we have obtained the potent acoustic representation, the next critical step is to align these with the visual features to create a unified multimodal comprehension approach. CLIP-based visual encoders naturally align visual encoding with text space, so we only need to apply projection layers to map the feature encoding of the audio modality into an embedding space that is compatible with LLMs for text generation.

Specifically, for every text caption that is paired with an audio modality, represented as (X_{text}, X_{audio}), the modality input undergoes a transformation to align with the text input embedding domain, resulting in the generation of Z_{audio}.

Formally, this alignment can be represented as:

$$p\left(X_{text}|X_{audio}\right) = \prod_{i=1}^{L} p_\theta \left(X_{text}^{[i]}|Z_{audio}, Z_{text}^{[1:i-1]}\right) \tag{6}$$

$$Z_{audio} = Projection_\theta\left(h_{latents}, g\left(X_{audio}\right)\right) \tag{7}$$

By aligning the audio and visual features in this manner, we ensure that the multimodal input is harmonized with the textual embedding domain. This allows us to infuse the rich environmental sound information into the LLM, thereby enhancing the overall quality and depth of the generated descriptions.

4. Experiments

4.1. Datasets

4.1.1. Training Datasets

MAD-v2. sMAD-v2(Soldan et al., 2022) is a large-scale dataset collected from Movie Audio Descriptions for the Language Grounding in Videos task. It comprises a total of 384K sentences grounded in over 1.2K hours of continuous videos from 650 different and diverse movies. MMAD exploits available audio descriptions of mainstream movies in MAD-v2 to train Movie Clip Contextual Alignment Module to align movie style visual token and AD.

JDIFLPS. This dataset(Xiao et al., 2017) contains 18,184 images, 8,432 identities, and 96,143 pedestrian bounding boxes, including character query and corresponding galleries for movies and various TV sitcoms. This dataset is used to train the abli matching detetiycted character visual token in keyframes to character images in character

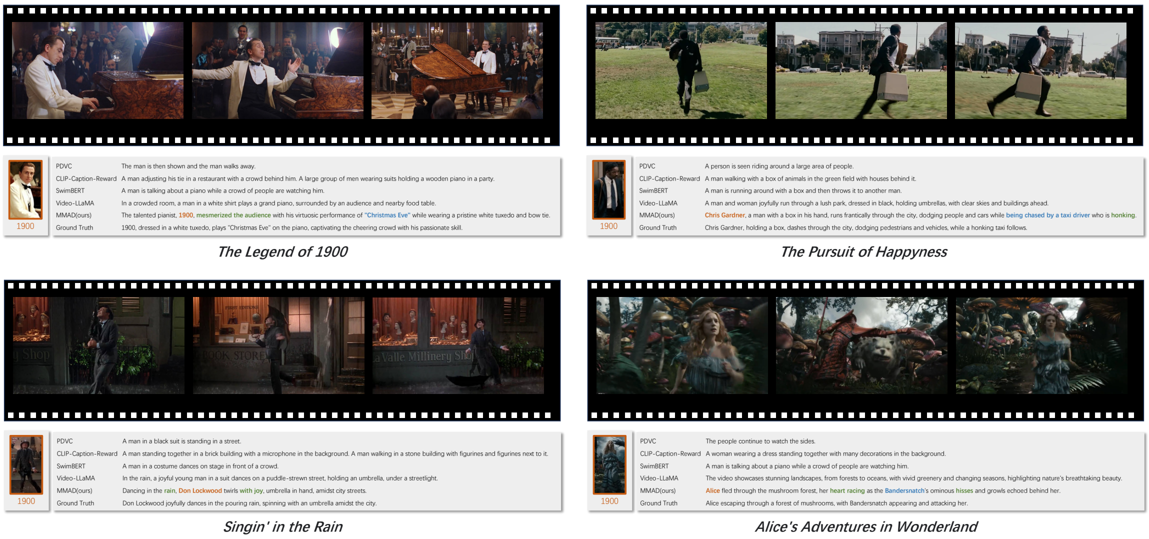

Figure 3: Qualitative results of our method. For a given movie clip, the Movie Clip Contextual Alignment Module determines the start and end timestamps for movie audio description generation, the Actor-tracking-aware Story Linking Module identifies the active character in the current movie clip, and the Audio-aware Feature Enhancing Module inputs the ambient music in the movie clip, which is used to assist in the generation of a more contextualized movie audio description.

list, enhancing Actor-tracking-aware Story Linking Module.

Clotho. Clotho(Drossos et al., 2020) consists of 4,981 diverse audio samples, each lasting 15 to 30 seconds, extracted from Freesound, supporting Audio-aware Feature Enhancing Module pretraining with a focus on diverse audio representations and 24,905 meticulously curated descriptions, emphasizing accurate caption of environmental acoustic representation.

AudioVault-AD. AudioVault-AD(Han et al., 2023) comprises over 3.3 million AD texts from 7,000+ films, emphasizing text quality over visual connections, ideal for standalone language model training, enriching movie-centric datasets, and surpassing competitors as the largest public AD textual collection by nearly tenfold. After aligning the multimodal inputs to the text space, the MMAD model outputs AD through LLM in a learnable prompt, and training the Movie Clip Contextual Alignment Module through AudioVault-AD helps to build informative and standards-compliant narration.

4.1.2. Testing Datasets

MovieNet. MovieNet(Huang et al., 2020) is a vast repository of visual data from 1,100 films, featuring 1.1 million annotated character frames and 42,000 scene demarcations, supporting diverse training objectives and reducing genre bias. This dataset is used to test the accuracy of aligning narration timestamps to video alignment actions in the Movie Clip Contextual Alignment Module.

M-VAD Names. M-VAD Names(Pini et al., 2019) contains the annotations of characters' visual appearances, in the form of tracks of face bounding boxes, and the associations with characters' textual mentions, when available. The released dataset contains more than 24k annotated video clips, including 63k visual tracks and 34k textual mentions, all associated with their character identities. This dataset is used to test the accuracy of character name matching in the Actor-tracking-aware Story Linking Module.

MC-eval. Existing benchmarks for evaluating movie audio description do not yet appear to contain datasets that simultaneously include movieglobal visual information, metadata for character information, previous AD and previous subtitles, and ambient music information. For MMAD model evaluation, we selected 10 renowned cinematic masterpieces, extracting over 20 segments from each, totaling 224 segments. Notably, 73% of these segments feature ambient music for more than two seconds. To each set of data in MC-eval, we added the overall screen global, character locator box, subtitles, audio description, and audio information. MC-eval is used as generated AD evaluation.

4.2. Implementation Details

The Audio-aware Feature Enhancing Module encodes ambient music features via ESResNeXt(Donahue et al., 2015) and aligns these features to the text domain. The Actor-tracking-aware Story Linking Module utilizes CLIP for character recognition and scene encoding. For voice transcription and identifying suitable audio description intervals in uncredited clips, we employ the

Movie Clip Contextual Alignment Module with WhisperX(Bain et al., 2023). These multimodal features feed into LLaMA2-70b(Touvron et al., 2023b), generating captions that resemble human-like language while keeping the parameters of LLaMA2 static to improve convergence and utilize its inherent reasoning abilities. During training, we use a batch size of 8, with each batch containing 10 movie clips (consisting of both subtitled and unsubtitled frames) and their corresponding audio descriptions. We also set the epoch to 20.

We use the Adam optimizer with an initial learning rate of 10^{-4}, decaying to 0, to independently adjust the learnable parameters in each module, all on eight 80G A100 GPUs. Given an average speaking rate of 180 words per minute, we crop movie clips to around 2 seconds each, limiting the character count for each audio description interval and each caption to 60 characters. We evaluate the models' character recognition using precision metrics like mean Average Precision (mAP) and Rank-1 (R1). For interval generation, we view it as multi-label and binary categorization, using accuracy, precision, and recall. For assessing audio description quality, we use BLEU-1, ROUGE-L and BertScore to measure word congruence with a reference and use RefCLIPScore to measure the similarity between the generated caption and the visual content, enhancing the generation of representative captions. Finally, we organized 10 vision health volunteers, 10 BVI people (including 3 totally blind and 7 partially sighted) for human evaluation via Likert scale(Joshi et al., 2015).

4.3. Quantitative Comparison

table 1 provides a comparative analysis of different video caption models, including our proposed MMAD framework, based on various input modalities (Visual - V, Audio - A, Language - L) and feature fusion approaches. The models compared include PDVC(Wang et al., 2021a), CLIP-Caption-Reward(Cho et al., 2022), SwinBERT(Lin et al., 2022), Video-LLaMA(Zhang et al., 2023), Vid2seq(Yang et al., 2023) and our MMAD framework.

In terms of metrics (B-1, R-L, BertS, RefCLIP-S, Human Evaluation), which serve as measures of caption quality, MMAD outperforms the other models. In objective evaluation metrics, MMAD's B-1 score of 44.5, R-L score of 39.2, BertS score of 60.6, and RefCLIP-S metric value of 0.825 are all higher than that of its closest competitor, SwimBERT.In human evaluation, MMAD's OA percentage of 72.8% is much higher than that of the Video-LLaMA's 59.2%.

4.4. Qualitative Comparison

fig. 3 presents the results of applying our MMAD system for generating movie audio descriptions on several films. The comparison between MMAD and other methods showcased in the figure vividly illustrates the superiority of our system. Specifically, MMAD is capable of producing more extensive and contextually rich descriptions.

This enhanced performance is largely attributable to the synergistic operation of the Audio-aware Feature Enhancing Module, the Actor-tracking-aware Story Linking Module, and the Movie Clip Contextual Alignment Module. These modules, by collectively leveraging the wealth of film information available, including character activity, ambient audio, and the optimal timing for descriptions, empower our system to generate highly detailed and contextually accurate movie audio descriptions.

4.5. Ablation Study

4.5.1. Effect of the Proposed Modules

In this part, we first study the influence of each proposed module and the employed LLM on the final Audio Description (AD) performance. We separately remove each module from our design and evaluate the resulting AD on the MC-eval dataset (table 2). The results show that removing the Audio-aware Feature Enhancing Module has minimal impact on the RefCLIP-S metrics, which primarily assess the correlation between movie frame visuals and text. However, removing the Actor-tracking-aware Story Linking and Movie Clip Contextual Alignment Modules, both crucial for visual information acquisition, significantly decreases the RefCLIP-S metrics and enlarges the gap between model-generated captions and the Ground Truth (GT). Furthermore, the LLM size significantly influences the AD quality, with LLaMA2-70b yielding more human-like captions than the 13b model, underscoring model complexity's impact.

4.5.2. Effect of Actor-tracking-aware Story Linking Module

The effectiveness of our Actor-tracking-aware Story Linking Module hinges greatly on the precision of character recognition. To evaluate this, we compare our proposed method with two face detection algorithms (RetinaFace(Deng et al., 2020b) and Abaw(Kollias, 2022))and two Re-Identification (ReID) algorithms (BoT(Luo et al., 2019) and LTReID(Wang et al., 2022)), widely used for accurately identifying main characters in movie datasets. The results are shown table 3. Comparing our approach with these five existing methods, our character recognition technique proves to significantly

Methods	Modality			Metric				Human Evaluation	
	V	A	L	B-1	R-L	BertS	RefCLIP-S	OA	CA
PDVC	✔	✗	✗	5.8	8.3	47.5	0.524	42.6%	✗
CLIP-Caption-Reward	✔	✗	✗	17.9	15.9	50.2	0.536	24.0%	✗
SwinBERT	✔	✗	✗	18.0	18.1	51.6	0.618	45.6%	✗
Video-LLaMA	✔	✗	✔	5.2	8.5	48.9	0.585	59.2%	✗
ours	✔	✔	✔	**44.5**	**39.2**	**60.6**	**0.825**	**72.8%**	✔

Table 1: The performance of the proposed MMAD framework and some video caption models with different input modalities (V-visual, A-audio, L-language) and feature fusion approaches in MC-eval dataset: we generated a comparison between movie audio description and GT based on a classical metric assessment of caption. MMAD takes into account all the input-able modalities, and achieves excellent caption results. In addition, MMAD has added human evaluation, which includes two indicators, "Overall information accessibility of the story (OA)" and "Character information accessibility (CA)", with the following statistical values Ratio of satisfied people/total number of researchers

Modules	B-1	R-L	BertS	RefCLIP-S
A2+A3+A4	18.5	13.2	39.9	0.682
A1+A3+A4	12.3	14.7	35.3	0.434
A1+A2+A4	19.5	14.9	38.1	0.311
A1+A2+A3+B4	28.9	16.3	43.8	0.582
A1+A2+A3+A4	**44.5**	**39.2**	**60.6**	**0.825**

Table 2: Ablation study on impact of proposed modules. A1 denotes Audio-aware Feature Enhancing Module, A2 denotes Actor-tracking-aware Story Linking Module, A3 denotes Movie Clip Contextual Alignment Module, A4 for LLaMA2-70b, B4 for LLaMA2-13b.

Methods	M-VAD Names		MC-eval(ours)	
	mAP	R1	mAP	R1
RetinaFace	35.2	44.9	32.4	42.8
Abaw	42.5	49.7	39.3	43.2
BoT	52.4	63.1	61.3	77.2
LTReID	55.8	62.9	61.7	78.8
Ours	**69.5**	**76.8**	**72.3**	**88.6**

Table 3: Ablation study of Actor-tracking-aware Story Linking Module on MC-eval M-VAD Names dataset.

Figure 4: Retrieval result visualization

enhance the performance. fig. 4 illustrates some examples of our recognition results.

4.5.3. Effect of Movie Clip Contextual Alignment Module

The Movie Clip Contextual Alignment Module plays a pivotal role in our design, as it integrates visual information from preceding dialog-rich clips into the caption generation process for the current clip. To understand the influence of this module, we explore the qualitative relationship between the quality of the generated descriptions and the number of frames considered from previous clips involving character dialogue. In this context, the quality of the Movie Audio Description is primarily evaluated using Bert-S and RefCLIP-S metrics. The relationship between the number of prior frames considered and the resulting description quality, as measured by these metrics, is depicted in fig. 5. We observe that as the number of prior subtitle-inclusive clips considered increases, the resulting movie audio description becomes more extensive. However, given the necessity to fit the narration within a specific time frame, a balance must be struck. When more than 96 frames are considered, the captions must be controlled for word count, leading to a more concise description. Consequently, this streamlining may result in a compromise in description quality.

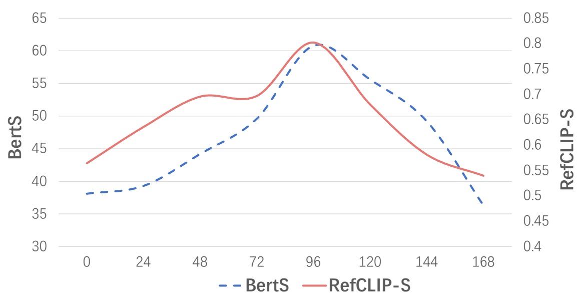

Figure 5: Ablation study on impact of number of frames considered from previous clips.

5. Conclusion and Future Work

This paper introduces the Multi-modal Movie Audio Description (MMAD), a novel framework for automated AD generation. Comprising three novel modules: the Audio-aware Feature Enhancing Module, the Actor-tracking-aware Story Linking Module, and the Subtitled Movie Clip Contextual Alignment Module. MMAD is designed to offer rich, extensive, and contextually aligned movie audio descriptions with the aid the large language models. Extensive experiments on established datasets have underscored the effectiveness of MMAD. However, it still faces some challenges: the character matching module based on pedestrian re-recognition can solve the problem of recognizing the same character under different lighting, but if the character changes clothes, it will have a greater impact on the accuracy, the design of a more robust character recognition module can help to realize a more specific and accurate caption; in addition, the current multimodal input is processed by modality In addition, the current multimodal input processing is realized by connecting projection layers through a modality encoder to map the modal information into the text embedding space that can be used in LLMs for caption generation. This mapping can hardly avoid the loss of input information, and proposing a model that realizes end-to-end, input raw data, and directly realizes the text output of LLMs is an important development direction in the future multi-modal accessible movie audio description field.

6. Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (Grant No.62372408).

7. Bibliographical References

Akhter Al Amin, Abraham Glasser, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2021. Preferences of deaf or hard of hearing users for live-tv caption appearance. In *International Conference on Human-Computer Interaction*, pages 189–201. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Virgínia P Campos, Luiz MG Gonçalves, Wesnydy L Ribeiro, Tiago MU Araújo, Thaís G Do Rego, Pedro HV Figueiredo, Suanny FS Vieira, Thiago FS Costa, Caio C Moraes, Alexandre CS Cruz, et al. 2023. Machine generation of audio description for blind and visually impaired people. *ACM Transactions on Accessible Computing*, 16(2):1–28.

Fuhai Chen, R. Ji, Jinsong Su, Yongjian Wu, and Yunsheng Wu. 2017. Structcap: Structured semantic embedding for image captioning. *Proceedings of the 25th ACM international conference on Multimedia*.

Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. 2018. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020a. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020b. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 374–378. IEEE.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 15–29. Springer.

Itamar Rocha Filho, Felipe Honorato, J Wallace Lucena, J Pedro Teixeira, and Tiago Maritan. 2021. An approach for automatic description of characters for blind people. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 53–56.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26.

Hongwei Ge, Zehang Yan, Kai Zhang, Mingde Zhao, and Liang Sun. 2019. Exploring overall contextual information for image captioning in human-like cognitive style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1754–1763.

Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10327–10336.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

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

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022.

Richang Hong, Meng Wang, Xiao-Tong Yuan, Mengdi Xu, Jianguo Jiang, Shuicheng Yan, and Tat-Seng Chua. 2011. Video accessibility enhancement for hearing-impaired users. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 7(1):1–19.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403.

JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible nuances: A caption system to visualize paralinguistic speech cues for deaf and hard-of-hearing individuals. pages 1–15.

A. Sophia Koepke, Andreea-Maria Oncescu, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2023. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 25:2675–2685.

Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50:171–184.

Dimitrios Kollias. 2022. Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2328–2336.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903.

Kuno Kurzhals, Fabian Göbel, Katrin Angerbauer, Michael Sedlmair, and Martin Raubal. 2020. A view on the viewer: Gaze-adaptive captions for videos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Siyuan Li, Li Sun, and Qingli Li. 2023a. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1405–1413.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Swinbert: End-to-end transformers with sparse attention for video captioning. In *CVPR*.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. pages 1487–1495.

Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D. Plumbley, and Wenwu Wang. 2022. Diverse audio captioning via adversarial training. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8882–8886.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*.

Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980.

Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. 2010. High five: Recognising human interactions in tv shows. pages 1–11.

Karol J Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021a. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Tanzila Rahman, Bicheng Xu, and Leonid Sigal. 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.

Rajib Rana. 2016. Gated recurrent unit (gru) for emotion classification from noisy speech. *arXiv preprint arXiv:1612.07778*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Andreas Robinson, Felix Jaremo Lawin, Martin Danelljan, Fahad Shahbaz Khan, and Michael Felsberg. 2020. Learning fast and robust target models for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7406–7415.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Mattia Soldan, A. Pardo, Juan Le'on Alc'azar, Fabian Caba Heilbron, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2021. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5016–5025.

Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. 2021. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862.

Anthony Teolis and John J Benedetto. 1998. *Computational signal processing with wavelets*, volume 182. Springer.

Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. 2016. Detecting text in natural image with connectionist text proposal network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 56–72. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: open and efficient foundation language models, 2023. *URL https://arxiv. org/abs/2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dirk Vander Wilt and Morwaread Mary Farbood. 2021. A new approach to creating and deploying audio description for live theater. *Personal and Ubiquitous Computing*, 25:771–781.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Anticipating visual representations from unlabeled video.

Pingyu Wang, Zhicheng Zhao, Fei Su, and Honying Meng. 2022. Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification. *IEEE Transactions on Multimedia*.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021a. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6847–6857.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*.

Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021b. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Wikipedia contributors. 2023. Audio description — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=Audio_description&oldid=1175592220`. [Online; accessed 13-October-2023].

Xian Wu, Guanbin Li, Qingxing Cao, Qingge Ji, and Liang Lin. 2018. Interpretable video captioning via trajectory structured localization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6829–6837.

Xuenan Xu, Heinrich Dinkel, Mengyue Wu, Zeyu Xie, and Kai Yu 0004. 2021. Investigating local and global information for automated audio captioning with transfer learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 905–909. IEEE.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*.

Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 444–454.

Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. 2010. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

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

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on* *computer vision and pattern recognition*, pages 5579–5588.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748.

8. Language Resource References

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

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

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Stefano Pini, Marcella Cornia, Federico Bolelli, Lorenzo Baraldi, and Rita Cucchiara. 2019. M-VAD Names: a Dataset for Video Captioning with Naming. *Multimedia Tools and Applications*.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5026–5035.

Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. 2017. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424.