LoCo-MAD: Long-Range Context-Enhanced Model Towards Plot-Centric Movie Audio Description

Jiayi Wang, Zihao Liu, and Xiaoyu Wu ✉

Communication University of China, Beijing 100024, China
{blindwang,liuzihao,wuxiaoyu}@cuc.edu.cn

Abstract. Movie Audio Description (MAD) aims to enable the visually impaired community to enjoy movies by transforming them into coherent and accurate audio descriptions. Due to the extended duration and complex plot natures of movies, MAD is in the early stages of research compared to other cross-modal text generation tasks. Current MAD methods fail to model long videos efficiently or integrate long-range context to generate plot-coherent descriptions. To address these challenges, we propose a Long-Range Context-Enhanced Movie Audio Description model (LoCo-MAD), which is trained in two stages. The first stage adapts an image-text pretrained model to a Pre-aligned Movie Encoder (PME), which utilizes learnable queries to obtain compact visual representations and is supervised by three multimodal objectives. The second stage builds LoCo-MAD with the pretrained PME, a Dynamic Selection Module (DSM), and a large language model. We project visual representations from PME into soft visual prompts and utilize DSM to select the most relevant descriptions and subtitles from a long range as contextual prompts. Then, a large language model integrates these multimodal prompts and generates plot-related movie descriptions. The proposed method is extensively evaluated on MAD-v2 and LSMDC datasets, where we achieve 23.7 and 20.0 CIDEr score, respectively. Our code will be released at `https://github.com/blindwang/LoCo-MAD`.

Keywords: Movie audio description · Multimodal learning · Large language model

1 Introduction

Movie audio description (MAD) aims to help the visually impaired community enjoy movies by automatically converting them into coherent and accurate audio narrations. It originated from the descriptive video service provided by television stations. Many streaming services have begun to offer this service, but they suffer from the high costs of manual production. Technically, MAD involves three steps: describing movie clips using natural language, converting textual descriptions into speech, and inserting speech at the appropriate position. This paper focuses on the first and most key step, which can be modeled as a cross-modal text generation task.

Due to the extended duration and complex plot natures of movies, MAD is in the early stages of research compared to other cross-modal text generation tasks (*e.g.* video captioning [23,30,37], dense video captioning [12,16,36], visual storytelling [15,20,39]). With the rise of LLM, many approaches [17,21,30] use LLM in video tasks to leverage common sense. However, it is hard to directly feed hour-long movies into LLM. Some related MAD approaches [10,11] use a simplified paradigm: inputting visual prompts of the current clip and supplementing it with contextual text prompts to generate descriptions. We address three-fold challenges within this paradigm: Firstly, there exists a modality gap between the LLM and the visual features of movies. Secondly, the model needs to acquire compact visual representations of the clip with dynamic duration to prompt the LLM. Finally, the complex plot makes long-range information crucial, where plenty of noise also exist. It is a challenge to increase the range of context information while reducing noise.

To advocate research in this direction, previous works explore reinforcement learning [14,33] and adversarial inference [27] to generate coherent sentences. However, these early works lack extensive common sense to understand the complex movie plot. Empowered by large language models (LLM) and vision-language pretrained models (VLP), AutoAD [11] and AutoAD-II [10] are capable of leveraging extensive common sense and contextual information. However, AutoAD lacks efficient visual modeling for long videos, and AutoAD-II cannot dynamically select relevant contextual information from a long range. Our model utilizes learnable queries to obtain compact visual features and propose a module to select relevant contextual information. MM-Narrator [40] is a GPT-4 empowered multimodal agent for MAD task. However, it cannot be fine-tuned to adapt to specific tasks, and the reliance on multiple large-scale pretrained models hinders its deployment. Additionally, the use of the closed-source GPT-4 limits its transparency and reproducibility. In contrast, our model can be trained end-to end.

In this paper, we propose a long-range context-enhanced model: LoCo-MAD. As shown in Fig. 1, our framework comprises two training stages. We pre-align the vision and language modalities and train the multimodal encoder to obtain compact visual features in the first stage. Then, we utilize a LLM to integrate the visual prompts projected from compact visual features and long-range contextual prompts in the second stage. The first stage targets at vision-language representation learning. We propose a Pre-aligned Movie Encoder (PME) module to align vision-language semantic representation. We employ a set of learnable queries to obtain compact visual representations, allowing computation-efficient training. To align the learnable queries with movie descriptions, we utilize three multimodal pretext tasks. The second stage targets at long-range context-enhanced description generation. It builds LoCo-MAD with the pretrained PME, a Dynamic Selection Module (DSM), and a large language model (LLM). DSM aims to select relevant information from extended and noisy contexts. DSM is supervised with a contextual contrastive loss (CCL), eliminating the need for additional labeled data. Instead of selecting from a fixed and short range of context,

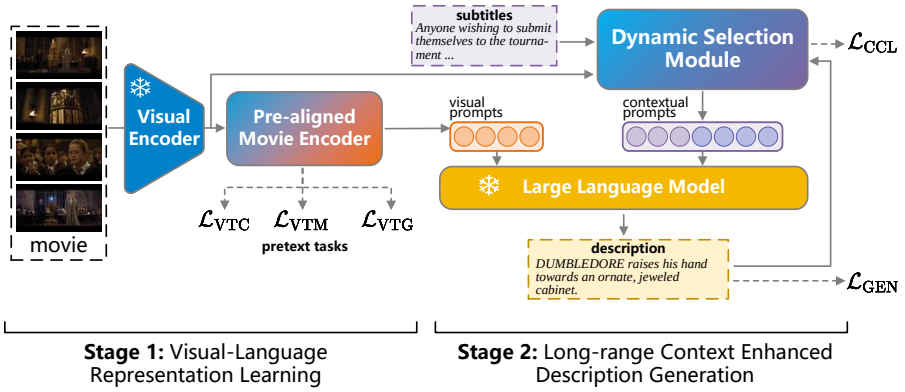

Fig. 1: Overview of LoCo-MAD's framework: LoCo-MAD is trained following a two-stage strategy. The first stage performs vision-language representation learning with PME and three pretext tasks. The second stage performs long-range context-enhanced description generation with DSM and LLM.

we dynamically select descriptions and subtitles from a long range, which serve as contextual prompts. We project visual representations from PME into soft visual prompts, and then we utilize a LLM to integrate these multimodal prompts and generate descriptions. After two training stages, LoCo-MAD can generate vision-grounded and plot-centric movie audio descriptions.

In summary, our contributions are as follows:

1) We present a two-stage training schema for MAD. We perform vision-language semantic alignment in the first stage and integrate long-range context for the generation of plot-centric movie descriptions in the second stage.
2) In the first stage, we propose PME to pre-align vision-language semantic features. We employ learnable queries to extract compact visual features, enabling computation-efficient training on long-form videos.
3) In the second stage, we propose DSM to select relevant information from extended and noisy contexts. Then, we utilize a LLM to integrate multimodal prompts and generate descriptions with extensive common sense.
4) Our framework is evaluated on MAD-v2 [11] and LSMDC [29]. The experimental results demonstrate that our framework exhibits the capabilities to generate accurate and plot-centric movie audio descriptions.

2 Related Work

2.1 Video-to-Text Generation

MAD belongs to the video-to-text generation task, which aims at generating textual descriptions for an input video. Related tasks include video captioning, dense video captioning, and visual storytelling. Video captioning aims to

generate a single sentence to describe the event in a short video. With the advent of VLP [17, 28], many works [23, 30, 37] have begun to use these advanced models for video captioning. Dense video captioning [12, 16] deals with untrimmed videos where we need to detect and describe multiple events. Recent works [4, 6, 7, 9, 36, 38] train localization and description module in an end-to-end manner. Visual storytelling [15, 20, 39] is to generate coherent sentences for a series of video clips or images. Previous works explore reinforcement learning [14, 33] and adversarial inference [27] to generate coherent sentences. In contrast, MAD takes untrimmed movies as input and generates descriptions that align with the plot development.

2.2 Vision-Language Pretrained Model

VLP models mainly learn the semantic correspondence between different modalities by pretraining on large-scale datasets, which has proved effective in improving the generalization capacity of video-to-text generation models [5]. CLIP [28] proposes the multimodal contrastive learning and lay the foundation for multimodal models to advance toward large-scale training. Many models [17, 18] begin to utilize multimodal contrastive learning for vision-language semantic alignment. To utilize the LLM and pretrained visual encoders, advanced VLP models [1, 17, 21] learn how to map visual features from pretrained visual encoders to the embedding space of the LLM. Given the impressive generalization capabilities of VLP models, we adapt an image-text pretrained model [17] to the PME and adopt multimodal contrastive learning for vision-language representation learning.

2.3 Movie Audio Description

MAD is challenging due to the dependency of contextual information and the complexity of visual elements in movies. AutoAD [11] leverages pretrained foundation models and bridges them with a simple Transformer-based mapping network, which cannot efficiently handle long videos. It incorporates a fixed range of contextual information via GPT-2 to generate plot-coherent descriptions. AutoAD-II [10] implements a flamingo-based [1] architecture for description generation. It handles the tasks of identifying who is in the scene and when to insert descriptions into the audio track for MAD, but it does not utilize long-range context to generate more consistent descriptions. MM-Narrator [40] is a training-free GPT-4 empowered multimodal agent. However, it relies on the closed-source GPT-4 and multiple large-scale pretrained models, which limits its transparency and reproducibility. To address existing problems, we introduce a two-stage framework for long-range context-enhanced movie description generation. In the first stage, we pre-align multimodal representations and utilize learnable queries to obtain compact visual features for computation-efficient training. In the second stage, we propose a dynamic selection module to select relevant contextual information from a long range as contextual prompts. Then,

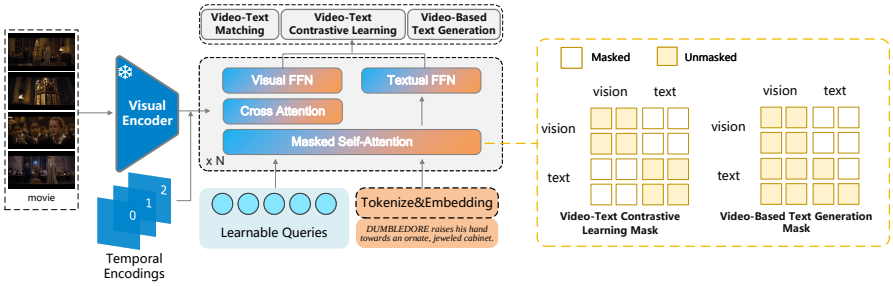

Fig. 2: Pre-aligned Movie Encoder. The learnable queries are utilized to obtain compact visual features. Different masked attention mechanisms are designed for each pretext task in the masked self-attention layer.

we integrate the visual prompts projected from the first stage and the selected contextual prompts via a LLM to generate plot-centric movie descriptions.

3 Method

Given a pre-segmented movie $\mathcal{M} = \{m_i\}_{i=1}^{n}$, the goal of MAD is to learn a function $\mathcal{F}_\theta$ with parameters θ that generates descriptions $\mathcal{T} = \{t_i\}_{i=1}^{n}$ for each segment m_i. Specifically, $\mathcal{F}_\theta$ generates the current description recurrently: $t_i = \mathcal{F}_\theta(m_{\leq i}, t_{<i})$, which is called the *recurrent* setting. For a more comprehensive evaluation of the performance, an *oracle* setting takes the ground truth $\hat{t}_{<i}$ as input instead the prediction: $t_i = \mathcal{F}_\theta(m_{\leq i}, \hat{t}_{<i})$. In this paper, we further formulate each clip m_i as a tuple of visual content v_i and subtitle s_i. The subtitle can be obtained from ASR or the movie producer. To generate both vision-grounded and plot-coherent movie audio descriptions, we train LoCo-MAD in two stages.

3.1 Stage 1: Vision-Language Representation Learning

In the first stage, we aim to align the semantic representation between vision and text modalities and obtain compact visual features, where we propose a PME and three pretext tasks.

Pre-aligned Movie Encoder. We utilize a pretrained visual encoder to extract the frame-level visual features $V \in \mathbb{R}^{L_v \times D_v}$ of current visual context v_i and a textual embedding layer to extract textual embeddings $T \in \mathbb{R}^{L_t \times D_t}$ of current description t_i.[1] L_v, L_t denote the length of video frames and text tokens, and D_v, D_t denote the dimensions of the visual encoder and textual embedding layer.

[1] For convenience, the subscript i, which denotes the current movie clip, will be omitted in the following paper.

The visual features of each frame are temporally encoded as V_t to distinguish between different frames. Considering the dynamic length of movie clips and the difficulty in learning temporal features from multiple frames, we employ unlearnable positional embeddings [34].

The proposed PME is shown in Fig. 2. The input consists of three parts: visual features V_t, learnable queries V_q, and textual embeddings T. Following the research of [17], the learnable queries are L_q learnable vectors used to obtain compact visual representations.

Firstly, the learnable queries and textual embeddings are fused in a shared multi-head self-attention layer:

$$z = [T; V_q] \ , \ F_z = \text{MultiHead}(z, z, z) \ , \tag{1}$$

where $\text{MultiHead}(q, k, v)$ denotes the multihead attention layer, and the symbol ';' denotes concatenation. Different masked attention mechanisms are designed for three pretext tasks to prevent information leakage. Next, the learnable queries extract compact visual representations from V_t through the cross-attention layer:

$$\tilde{F}_v = \text{MultiHead}(F_z, V_t, V_t) \ . \tag{2}$$

Finally, two feed-forward networks (FFN) process the learnable queries and textual features respectively to enhance the representation learning:

$$\tilde{q} = \text{FFN}(F_z) \ , \ \tilde{t} = \text{FFN}(\tilde{F}_v) \ , \tag{3}$$

where FFN denotes the feed-forward network, $\tilde{q}$ is the compact visual features and $\tilde{t}$ is the textual features with $\tilde{t}_0$ as the output feature of the `[CLS]` token.

Multimodal Pretext Tasks. Inspired by [17], we utilize three pretext tasks to bridge the semantic gap between the vision and language modalities. **Video-Text Contrastive Learning (VTC)** aims to align the vision and text by bringing the paired semantic representations closer. It is achieved by contrasting the video-text similarity of a positive pair against those of negative pairs. For each video-text pair, the similarity between each query $\tilde{q}$ and the textual features $\tilde{t}_0$ from the output of PME is computed, and we use the highest one as the similarity of each pair. We compute the InfoNCE loss [26] between vision and text modalities to perform contrastive learning. For VTC task, vision and text modalities cannot exert attention on each other in the masked self-attention layer of PME, as shown in Fig. 2. **Video-Text Matching (VTM)** aims to learn fine-grained representation between vision and text. VTM is a binary classification task to evaluate whether the description matches the movie clip. We feed $\tilde{q}$ into a binary classifier to get the probabilities of all queries, and then average them to obtain the final matching score. We employ the binary cross-entropy loss to train, where we use the hard negative mining [19] method when performing the selection of negative samples. For the VTM task, both vision and language modalities need to attend to each other so that we do not use the masks during encoding. **Video-based Text Generation (VTG)** aims

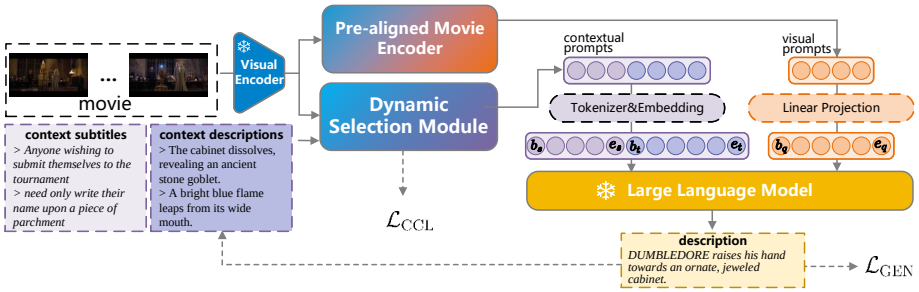

Fig. 3: Overview of the long-range context-enhanced generation learning stage (stage2). A dynamic selection module is used to select relevant descriptions and subtitles from context proposal, which serve as contextual prompts. Then, a LLM is used to integrate visual prompts from the PME and contextual prompts. b_* and e_* are special tokens to differentiate visual prompts from contextual prompts.

to obtain aligned multimodal representation by guiding the PME to generate vision-conditioned descriptions. VTG generates text in an auto-regressive manner, and we use a causal mask during encoding. We use the cross-entropy loss with label smoothing [32] to train. Our model is better equipped to understand the relationship between visual content and textual descriptions after the first stage training. With three pretext tasks combined, the total loss of this stage is defined as follows:

$$\mathcal{L}_{s1} = \mathcal{L}_{\mathrm{VTC}} + \mathcal{L}_{\mathrm{VTM}} + \mathcal{L}_{\mathrm{VTG}} \ . \tag{4}$$

3.2 Stage 2: Long-Range Context-Enhanced Generation Learning

In this stage, we aim to generate long-range context-enhanced descriptions. As shown in Fig. 3, we project the compact visual representations obtained from the PME into soft visual prompts and propose a DSM to select relevant contextual descriptions and subtitles from a long range as contextual prompts. Then, we utilize a LLM to generate descriptions conditioned on the visual and contextual prompts.

Dynamic Selection Module. DSM aims to select the top k relevant descriptions and subtitles as contextual prompts for the current movie clip. The input of DSM includes the current movie clip and the context proposal. The context proposal comprises descriptions $t_{<i}$ and subtitles $s_{\leq i}$ from previous and current movie clips. We select relevant contexts according to the similarities between the context descriptions and the current movie clip. The corresponding subtitles for the context description are selected simultaneously. Context descriptions and subtitles can provide auxiliary plot information for the coherent description generation in this stage. Under the *recurrent* setting, we utilize the generated movie descriptions of our model as context proposal as shown in Fig. 3.

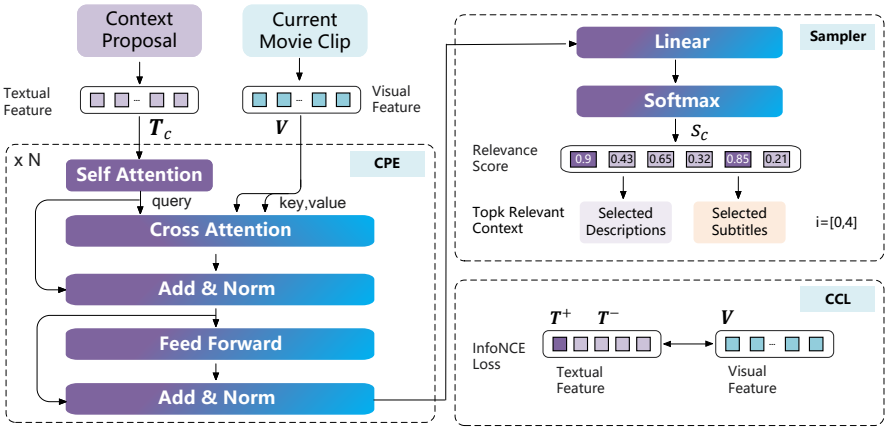

Fig. 4: Dynamic Selection Module. The Context Proposal Encoder (CPE) encodes multimodal features of vision and context descriptions. The Sampler computes relevance scores based on multimodal features. A Contextual Contrastive Loss (CCL) is utilized to distinguish the positive context-video pairs from negative pairs.

As shown in Fig. 4, DSM consists of three components: a Context Proposal Encoder (CPE), a Sampler, and a Contextual Contrastive Loss (CCL).

Context Proposal Encoder (CPE). The input of CPE includes textual features of the context proposal $T_c \in \mathbb{R}^{L_c \times D_c}$ and visual features V_t extracted in the first stage. T_c is extracted from a pretrained textual and visual encoder. L_c denotes the length of context proposal and D_c denotes the dimension of the textual encoder. T_c are input into the self-attention layer to obtain the mutual attention between them:

$$T_{sa} = \text{MultiHead}(T_c, T_c, T_c) \ . \tag{5}$$

To obtain the relevance of the current movie clip V within the context proposal T_c, we utilize cross-attention where textual features serve as the queries, and the visual features serve as keys and values. This process results in vision-conditioned textual features T_{ca}, which represent the correspondence between the two modalities. We apply the cross-attention layer as follows:

$$T_{ca} = \text{MultiHead}(T_{sa}, V, V) \ . \tag{6}$$

Sampler. The Sampler contains a linear layer $\text{Proj}_d(\cdot)$ and an activation function $\text{softmax}(\cdot)$ to obtain relevance scores for each context description:

$$s_c = \text{softmax}(\text{Proj}_d(T_{ca})) \ . \tag{7}$$

Based on the relevance scores s_c, we sample top k relevant descriptions and subtitles from context proposal, which can be denoted as $t^* = \{t_i^*\}_{i=1}^k$ and $s^* = \{s_i^*\}_{i=1}^k$.

Contextual Contrastive Loss (CCL). The CCL guides DSM to identify the descriptions related with the current movie clip. As shown in Fig. 4, the positive context T^+ is defined by averaging the textual features $\mathcal{K}$ of t^* with relevance scores as weights, while negative context T^- consists of the textual features $\mathcal{R}$ of remaining descriptions in context proposal:

$$T^+ = \frac{1}{k}\sum_{T_i \in \mathcal{K}} s_c T_i \ , \ T^- = [T_i]_{T_i \in \mathcal{R}} \ . \tag{8}$$

The positive context-video pairs (T^+, V) and negative pairs (T^-, V) are utilized for contrastive learning. We apply the contextual contrastive loss [26] to distinguish the positive context-video pairs from negative pairs:

$$\mathcal{L}_{\text{CCL}} = -\log \frac{\exp(\text{sim}(T^+, V)/\tau)}{\sum_{t \in [T^+, T^-]} \exp(\text{sim}(t, V)/\tau)} \ , \tag{9}$$

where τ denotes the temperature coefficient, and $\text{sim}(\cdot)$ represents the cosine similarity function.

Language Decoder. The inputs of language decoder are the visual prompts projected from visual representations $\tilde{q}$ which are obtained from the pretrained PME, and contextual prompts which include the selected context descriptions t^* and subtitles s^* from the DSM. We add special tokens b_t,e_t and b_s,e_s to distinguish subtitles from descriptions when concatenating them. Then we feed concatenated context descriptions and subtitles into the embedding layer to get the textual features:

$$c = [b_t; t_i^*; e_t; b_s; s_i^*; e_s]_{i=1}^{k} \ , \ \overline{c} = \text{Embedding}(\text{Tokenizer}(c)) \ . \tag{10}$$

The visual representations $\tilde{q}$ are linearly mapped to the feature dimensions accepted by LLM:

$$\tilde{q}' = \text{Proj}_l(\tilde{q}) \ , \ \overline{q} = [b_q; \tilde{q}'; e_q] \ , \tag{11}$$

where b_q and e_q are special tokens to differentiate visual prompts from contextual text. We utilize LLM to integrate visual and contextual prompts, thereby enhancing the generation of plot-centric and high-quality descriptions:

$$t = \text{LLM}([\overline{q}; \overline{c}]) \ , \tag{12}$$

where t denotes the generated movie description for the current movie clip. We utilize cross-entropy loss between ground truth and predicted description as the training objective of this stage, which can be denoted as $\mathcal{L}_{\text{GEN}}$.

Finally, the total loss of this stage is defined as follows:

$$\mathcal{L}_{s2} = \frac{1}{2}(\mathcal{L}_{\text{CCL}} + \mathcal{L}_{\text{GEN}}) \ . \tag{13}$$

4 Experiments

4.1 Experimental Setup

Datasets.

MAD-v2. MAD-v2 dataset is established in AutoAD [11], which comprises a total of 264K descriptions grounded in 892 hours of continuous videos from 488 different movies. It also provides subtitles in movies. MAD-eval is compromised of 6.5K descriptions and 10.6K subtitles from 10 movies. Moreover, MAD-v2 releases named and unnamed version for both training and testing dataset.

LSMDC. The task of the Multi-Sentence Description of LSMDC 2019 [29] is to generate consecutive captions for a given set of five clips. The dataset contains over 128K clips from 200 movies and has four splits: 20,283 training, 1,486 validation, 2,018 public test, and 1,923 blind test samples. Our performance is evaluated on the public test set, considering the blind test is not available.

Architecture. For the **visual encoder**, we use the CLIP ViT-L/14 model [28], which is a 24-layer transformer encoder that outputs 1×768 feature vectors for each input frame. The MAD dataset provides these features, while for the LSMDC dataset, we extract the visual features for the provided video clips. We perform frame sampling at 5 frames per second (FPS). For the **PME**, we use 12 transformer blocks with 32 learnable queries. We initialize PME with the pretrained weights of BLIP-2 ViT-L. For the **DSM**, we use 2 transformer blocks. For the MAD dataset, we set the length of context proposal to 10. The LSMDC dataset provides five consecutive video clips, so the context range is fixed to 5. For the textual feature in the DSM, we utilize the language encoder of CLIP ViT-L/14 model [28], which outputs 1×768 feature vectors for each sentence. For the **language decoder**, we use OPT-2.7B [41]. We limit the generated number of tokens to 36, since most movie audio descriptions are less than 36 tokens. We truncate the concatenated descriptions and subtitles to 200 tokens. In the ablation experiments, we use OPT-350M.

Evaluation Metrics. We use classic metrics including ROUGE-L [22] (R-L), CIDEr [35] (C), SPICE [2] (S), METEOR [8] (M), and BertScore [42] (BertS). We also report a recall-based metric *Recall@k within Neighbours* (R@k/N) proposed in AutoAD-II [10]. To compute recall, the text similarity is measured by BertScore, which evaluates word matching between a candidate sentence and reference sentence with pretrained BERT embeddings. A higher value indicates better text generation for all metrics.

4.2 Quantitative Results

Results on the MAD task. In Tab. 1, we compare our method with existing visual captioning and MAD methods on MAD-v2 dataset. LoCo-MAD achieves

Table 1: Results on the MAD task. Models are evaluated on the MAD-eval-Named benchmark. We report the results evaluated under different settings, including the *local* (without context), *oracle* (with ground truth descriptions as context), and *recurrent* (with previously predicted descriptions as context). WV refers to Web-Vid2M dataset. AV refers to AudioVault-AD dataset. MN refers to MovieNet dataset. S1 and S2 refers to Stage1 and Stage2 of LoCo-MAD.

Model	LLM	Setting	Pretrain Data	R-L	C	S	R@5/16
ClipCap [24]	GPT-2(1.5B)	Local	CC3M [31]	8.5	4.4	1.1	36.5
CapDec [25]	GPT-2(1.5B)	Local	AV [11]	8.2	6.7	1.4	-
LoCo-MAD: S1	-	Local	-	**11.1**	11.6	**3.4**	49.4
LoCo-MAD: S2	OPT(2.7B)	Local	-	10.9	**12.2**	3.3	**50.1**
AutoAD [11]	GPT-2(1.5B)	Oracle	AV&WV [3]	13.9	21.9	4.8	-
LoCo-MAD: S2	OPT(1.3B)	Oracle	-	14.0	22.5	6.0	**47.1**
LoCo-MAD: S2	OPT(2.7B)	Oracle	-	**14.1**	**23.7**	**6.3**	45.7
AutoAD [11]	GPT-2(1.5B)	Recurrent	AV&WV	11.9	14.3	4.4	42.1
AutoAD-II [10]	GPT-2(1.5B)	Recurrent	AV&WV&MN [13]	13.4	**19.5**	-	**50.8**
MM-Narrator [40]	GPT-4	Recurrent	-	12.1	11.6	4.5	48.0
LoCo-MAD: S2	OPT(2.7B)	Recurrent	-	**13.9**	18.8	**5.5**	43.9

Table 2: Results on the MAD-eval-Unnamed dataset. The first stage of LoCo-MAD is evaluated under the *local* setting, while other methods are under the *oracle* setting. All models are trained on the MAD dataset's Named and Unnamed versions to evaluate the cross-domain testing results.

Model	Train set	MAD-eval-Unnamed			
		R-L	C	S	BertS
AutoAD	Unnamed	15.9	14.5	10.5	26.7
	Named	11.4	10.0	3.1	22.5
LoCo-MAD (Stage1)	Unnamed	**16.9**	17.1	10.1	**47.3**
	Named	16.8	16.4	10.1	47.1
LoCo-MAD (Stage2)	Unnamed	16.2	**19.8**	**11.4**	46.8
	Named	13.2	17.5	5.4	45.7

state-of-the-art performance on C scores under the *local* (without context) and *oracle* (with ground truth descriptions as context) settings. AutoAD-II is additionally trained on the character name labels from MovieNet [13] and our method performs worse than it (C 18.8 vs 19.5) under the *recurrent* setting (with previously predicted descriptions as context). Together with its competitive performance on other metrics, LoCo-MAD demonstrates the effectiveness of two-stage training schema. Specifically, it is unnecessary for LoCo-MAD to train on pretraining dataset, *e.g.* AudioVault-AD [11] and Web-Vid2M [3]. LoCo-MAD achieves more competitive performance (C 22.5 vs 21.9) under the *oracle* setting using the same size of LLM. In Tab. 2, we report the result evaluated on the MAD-eval-Unnamed dataset. The model is trained on the MAD dataset's Named and Unnamed versions to evaluate cross-domain testing results. Our method achieves higher scores across all metrics compared to AutoAD. Addi-

Table 3: Results on the public test set of LSMDC.

Model	Training Data	C	M
Official Baseline [27]	CC3M	11.9	8.3
TAPM [39]	LSMDC	15.4	8.4
AutoAD	MAD-v2-Unnamed & LSMDC	17.5	7.5
LoCo-MAD	LSMDC	**20.0**	**9.5**

tionally, our method maintain an acceptable performance in the cross-domain testing.

Results on the multi-sentence description task. In Tab. 3, we compare our method with existing visual storytelling and MAD methods on LSMDC 2019 dataset, in which the model generates five corresponding descriptions for five consecutive clips. AutoAD and our method take previous predicted descriptions as context, while other methods take all five clip together for description generation. Compared to AutoAD, our approach achieves 14.3% (C 20.0 vs 17.5) and 26.7% (M 9.5 vs 7.5) growth in the C and M metrics, respectively.

Table 4: Ablations of two-stage training on MAD-eval-Named.

Method	R-L	C	S	BertS
Stage1	11.1	11.6	3.4	44.0
Stage2	13.0	18.2	5.4	45.2
Stage1+Stage2	**14.0**	**22.5**	**6.0**	**46.7**

Table 5: Ablations of three pretext tasks on MAD-eval-Unnamed.

Loss Function	R-L	C	S	BertS
Combined loss	**16.9**	**17.1**	10.1	**47.3**
w/o VTC loss	16.6	16.0	10.0	47.1
w/o VTM loss	16.6	15.8	**10.3**	47.1

Ablation Studies.

Effect of two-stage training We validate the effectiveness of two-stage training schema via ablation experiments. As shown in Tab. 4, our model achieve the best performance with combined two-stage training, which demonstrates the effectiveness of two-stage training schema.

Effect of PME We validate the effectiveness of three pretext tasks in the PME with a cumulative ablation in Tab. 5. The goal of pretext tasks is to align the semantic representation between the vision and language modalities. The alignment performance of PME affects the visual prompts provided to the language model, which in turn impacts its ability to generate movie descriptions. The unified loss demonstrates superior performance across all evaluation metrics, notably achieving a C score of 17.1. This performance can be attributed to the fine-grained alignment between vision and language modalities in the PME. With the removal of the VTC loss and VTM loss, there is a decline in the C score of approximately 4.3% (C 16.0 vs 17.1) and 1.3% (C 15.8 vs 16.0), respectively. This proves the efficacy of multimodal alignment loss (VTC and VTM) in the representation learning stage.

Table 6: Ablations of the DSM and context length on MAD-eval-Unnamed dataset. The performance is evaluated by C score. For model w/ DSM, context length means the selected top k relevant context.

Method	Context Length			
	2	3	4	6
AutoAD	13.5	17.0	-	19.5
w/ DSM	**20.5**	**23.7**	**20.5**	**21.9**
w/o DSM	17.2	18.8	20.2	19.1

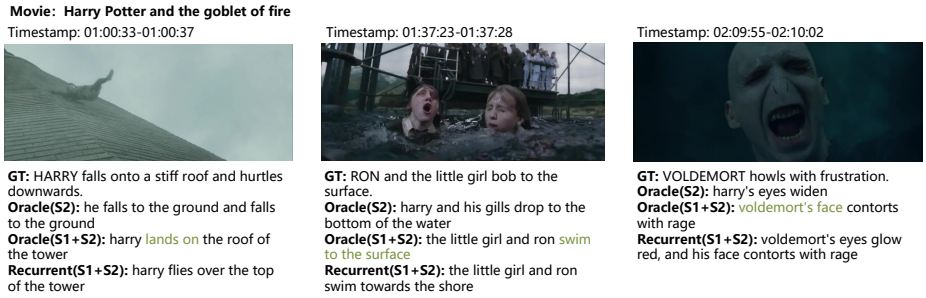

Fig. 5: Qualitative examples of generated movie descriptions. We highlight predictions under both the *oracle* and *recurrent* settings.

Effect of DSM and context length. Our model w/ DSM consistently outperforms AutoAD and model w/o DSM across various numbers of context length, as shown in Tab. 6. Through the design of DSM, our model achieves better performance (C 23.7 vs 19.5) with shorter but plot-related contexts. With fewer context prompts inputted into the LLM, the model runs fewer times, leading to an overall speedup. Therefore, our model strikes a favorable balance between efficiency and performance. Our proposed model with DSM shows performance degradation with longer context lengths, while AutoAD and the model without DSM exhibit a similar performance trend. For the model w/ DSM, longer contexts introduce more noise from the extended context proposals (set to 10), making it difficult for the LLM to integrate the information effectively, which leads to performance degradation.

4.3 Qualitative Results

Fig. 5 shows three qualitative examples of our model. It shows that the LoCo-MAD is able to generate vision-grounded and plot-coherent descriptions under both *oracle* and *recurrent* settings. Without the first stage of multimodal alignment, the model suffers from inaccurate character identification and action recognition. This verifies the effectiveness of two-stage training. Fig. 6 shows two qualitative examples of selected relevant contexts from DSM. The DSM selects two relevant descriptions from the context proposal for each movie clip. Without

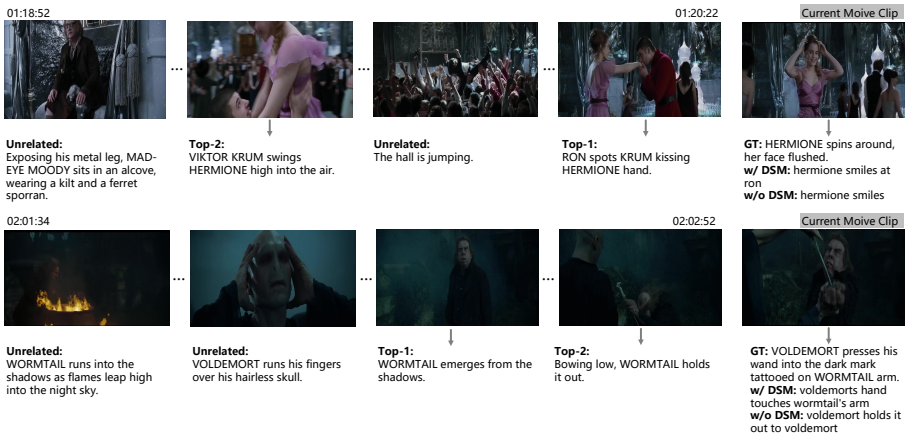

Fig. 6: Qualitative examples of selected relevant context from DSM. The number of context proposal is set to 15. Our models are evaluated under the *oracle* setting.

the DSM, our model fails to associate relevant plots in the previous movie clips. It proves that DSM is able to guide our model to understand the complex plots and help generate plot-centric descriptions.

5 Conclusion

This paper focuses on the generation of plot-centric movie audio descriptions. We propose a two-stage training framework for vision-language representation learning and long-range context-enhanced generation learning. We demonstrate the effectiveness of the PME and DSM in generating the vision-grounded and plot-centric movie descriptions. However, LoCo-MAD has limitations in identifying movie characters and relying on the annotated AD timestamps to segment movies. We will tackle these problems in the future work.

Acknowledgements. This work was supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0900701 and 2021YFF0602103, and in part by Natural Science Foundation of China (No.61801441). We also thank the research funds under the High-quality and Cutting-edge Disciplines Construction Project for Universities in Beijing (Internet Information, Communication University of China).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A.,

Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022. vol. 35, pp. 23716–23736. New Orleans, LA, USA (2022)

2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision(ECCV). vol. 9909, pp. 382–398. Springer (2016)
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 1708–1718. IEEE, Montreal, QC, Canada (2021). https://doi.org/10.1109/ICCV48922.2021.00175
4. Chadha, A., Arora, G., Kaloty, N.: iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. ArXiv **abs/2011.07735** (2020)
5. Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., Xu, B.: VLP: A survey on vision-language pre-training. Int. J. Autom. Comput. **20**(1), 38–56 (2023). https://doi.org/10.1007/S11633-022-1369-5
6. Chen, S., Jiang, Y.: Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 8425–8435. Computer Vision Foundation / IEEE, Virtual Event (2021). https://doi.org/10.1109/CVPR46437.2021.00832
7. Deng, C., Chen, S., Chen, D., He, Y., Wu, Q.: Sketch, ground, and refine: Top-down dense video captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. pp. 234–243. Computer Vision Foundation / IEEE, Virtual Event (2021). https://doi.org/10.1109/CVPR46437.2021.00030
8. Denkowski, M.J., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. pp. 376–380. ACL, Baltimore, USA (2014)
9. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., Huang, J.: Weakly supervised dense event captioning in videos. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018. pp. 3063–3073. Montréal, Canada (2018)
10. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad II: the sequel - who, when, and what in movie audio description. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023. pp. 13599–13609. IEEE, Paris, France (2023). https://doi.org/10.1109/ICCV51070.2023.01255
11. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad: Movie description in context. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 18930–18940. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01815
12. Huang, G., Pang, B., Zhu, Z., Rivera, C., Soricut, R.: Multimodal pretraining for dense video captioning. In: Wong, K., Knight, K., Wu, H. (eds.) Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020. pp. 470–490. Association for Computational Linguistics, Suzhou, China (2020)

13. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: Movienet: A holistic dataset for movie understanding. In: Computer Vision - ECCV 2020 - 16th European Conference. Lecture Notes in Computer Science, vol. 12349, pp. 709–727. Springer, Glasgow, UK (2020). https://doi.org/10.1007/978-3-030-58548-8_41
14. Huang, Q., Gan, Z., Celikyilmaz, A., Wu, D.O., Wang, J., He, X.: Hierarchically structured reinforcement learning for topically coherent visual story generation. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. pp. 8465–8472. AAAI Press, Honolulu, Hawaii, USA (2019). https://doi.org/10.1609/AAAI.V33I01.33018465
15. Huang, T.H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D.: Visual storytelling. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. pp. 1233–1239. ACL, San Diego California, USA (2016)
16. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: IEEE International Conference on Computer Vision, ICCV 2017. pp. 706–715. IEEE Computer Society, Venice, Italy (2017). https://doi.org/10.1109/ICCV.2017.83, https://doi.org/10.1109/ICCV.2017.83
17. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, ICML 2023. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR, Honolulu, Hawaii, USA (2023)
18. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, ICML 2022. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR, Baltimore, Maryland, USA (2022)
19. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: Advances in Neural Information Processing Systems(NIPS). vol. 34, pp. 9694–9705. Curran Associates, Inc., Virtual Event (2021)
20. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Video storytelling: Textual summaries for events. IEEE Trans. Multim. **22**(2), 554–565 (2020). https://doi.org/10.1109/TMM.2019.2930041
21. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. ArXiv **abs/2311.10122** (2023)
22. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Annual Meeting of the Association for Computational Linguistics(ACL). pp. 74–81. ACL, Barcelona, Spain (2004)
23. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Chen, X., Zhou, M.: Univilm: A unified video and language pre-training model for multimodal understanding and generation. ArXiv **abs/2002.06353** (2020)
24. Mokady, R., Hertz, A.: Clipcap: Clip prefix for image captioning. ArXiv **abs/2111.09734** (2021)
25. Nukrai, D., Mokady, R., Globerson, A.: Text-only training for image captioning using noise-injected CLIP. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 4055–4063. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/V1/2022.FINDINGS-EMNLP.299

26. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. ArXiv **abs/1807.03748** (2018)
27. Park, J.S., Rohrbach, M., Darrell, T., Rohrbach, A.: Adversarial inference for multi-sentence video description. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. pp. 6598–6608. Computer Vision Foundation / IEEE, Long Beach, CA, USA (2019). https://doi.org/10.1109/CVPR.2019.00676
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning(ICML). vol. 139, pp. 8748–8763. PMLR, Virtual Event (2021)
29. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C.J., Larochelle, H., Courville, A.C., Schiele, B.: Movie description. Int. J. Comput. Vis. **123**(1), 94–120 (2017). https://doi.org/10.1007/S11263-016-0987-1
30. Seo, P.H., Nagrani, A., Arnab, A., Schmid, C.: End-to-end generative pretraining for multimodal video captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022. pp. 17959–17968. IEEE, New Orleans, LA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01743
31. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018. pp. 2556–2565. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/V1/P18-1238
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. pp. 2818–2826. IEEE Computer Society, Las Vegas, NV, USA (2016). https://doi.org/10.1109/CVPR.2016.308
33. Tang, J., Wang, J., Li, Z., Fu, J., Mei, T.: Show, reward, and tell: Adversarial visual story generation. ACM Trans. Multim. Comput. Commun. Appl. **15**(2s), 54:1–54:20 (2019). https://doi.org/10.1145/3291925
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.U., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems(NIPS). vol. 30, pp. 5998–6008. Long Beach, CA (2017)
35. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575. IEEE, Boston, MA, USA (2015)
36. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. pp. 6827–6837. IEEE, Montreal, BC, Canada (2021). https://doi.org/10.1109/ICCV48922.2021.00677
37. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., Zhou, J.: mplug-2: A modularized multi-modal foundation model across text, image and video. In: International Conference on Machine Learning, ICML 2023. vol. 202, pp. 38728–38748. PMLR, Honolulu, Hawaii, USA (2023)
38. Yang, A., Nagrani, A., Seo, P.H., Miech, A., Pont-Tuset, J., Laptev, I., Sivic, J., Schmid, C.: Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023. pp. 10714–10726. IEEE, Vancouver, BC, Canada (2023). https://doi.org/10.1109/CVPR52729.2023.01032

39. Yu, Y., Chung, J., Yun, H., Kim, J., Kim, G.: Transitional adaptation of pretrained models for visual storytelling. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 12658–12668. Computer Vision Foundation / IEEE, virtual (2021). https://doi.org/10.1109/CVPR46437.2021.01247
40. Zhang, C., Lin, K.Q., Yang, Z., Wang, J., Li, L., Lin, C.C., Liu, Z., Wang, L.: Mmnarrator: Narrating long-form videos with multimodal in-context learning. ArXiv **abs/2311.17435** (2023)
41. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M.T., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: Opt: Open pre-trained transformer language models. ArXiv **abs/2205.01068** (2022)
42. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations, ICLR 2020. OpenReview.net, Addis Ababa, Ethiopia (2020)