# Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies

**Yingqiang Gao**[UZH], **Lukas Fischer**[UZH], **Alexa Lintner**[zhaw], **Sarah Ebling**[UZH]

[UZH]Department of Computational Linguistics, University of Zurich, Switzerland
{yingqiang.gao, fischerl, ebling}@cl.uzh.ch

[zhaw]School of Applied Linguistics, Zurich University of Applied Sciences, Switzerland
alexa.lintner@zhaw.ch

## Abstract

Audio descriptions (ADs) function as acoustic commentaries designed to assist blind persons and persons with visual impairments in accessing digital media content on television and in movies, among other settings. As an accessibility service typically provided by trained AD professionals, the generation of ADs demands significant human effort, making the process both time-consuming and costly. Recent advancements in natural language processing (NLP) and computer vision (CV), particularly in large language models (LLMs) and vision-language models (VLMs), have allowed for getting a step closer to automatic AD generation. This paper reviews the technologies pertinent to AD generation in the era of LLMs and VLMs: we discuss how state-of-the-art NLP and CV technologies can be applied to generate ADs and identify essential research directions for the future.

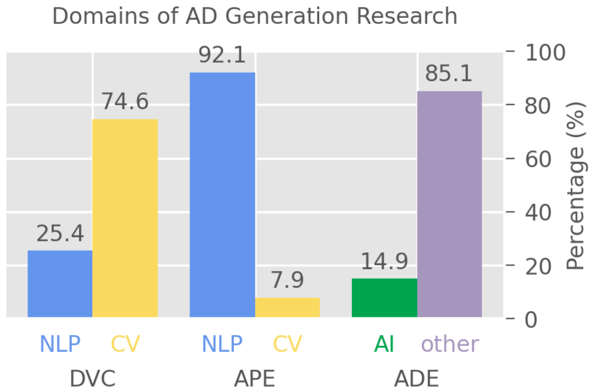


Figure 1: Domain contributions of AD-generation-related publications reviewed in this survey. DVC, APE, and ADE represent the three main steps of AD generation systems: Dense Video Captioning, AD Post-Editing, and Audio Description Evaluation, respectively. The figure illustrates the varying contributions to AD generation research across different domains. For ADE, "other" represents non-AI-related research disciplines such as psychology.

## 1 Introduction

### 1.1 Background

The formalization of AD as a public service can be traced back only to the early 1980s in the United States (Mazur, 2020). Initially introduced in the theater, AD services have expanded to a wide range of contexts, including television programs, movies, art galleries, and museums, in order to mitigate the information loss experienced by blind individuals and individuals with visual impairments. A significant milestone in the development of ADs was achieved in 2010, when the European Parliament included the provision of "accessible audiovisual media services" in its directives for that year (Reviers, 2016). Since then, AD research has garnered widespread interest.

ADs are traditionally produced by professional audio describers. The production process begins with acquiring the broadcast material, ideally complemented with time codes. Audio describers then review the material and create AD scripts (ideally in cooperation with blind audio describers or audio describers with visual impairments) tailored to the broadcast content. The final step involves recording the AD scripts in a studio, potentially with the involvement of a blind prooflistener, or synthesizing the speech and subsequently mixing the acoustic ADs with the original broadcast audio (Fryer, 2016). Producing ADs for a 90-minute movie can take approximately 35 to 40 working hours[1], underscoring the vast amount of information that remains inaccessible to blind and visually impaired individuals without these ADs. This highlights the significant value and necessity of the work carried out by professional audio describers.

However, training professional audio describers is a time-intensive process (Matamala and Orero, 2007; Jankowska, 2017; Colmenero et al., 2019;

[1]Experience shared by audio describers we work with.



*Findings of the Association for Computational Linguistics: NAACL 2025*, pages 471–490
April 29 - May 4, 2025 

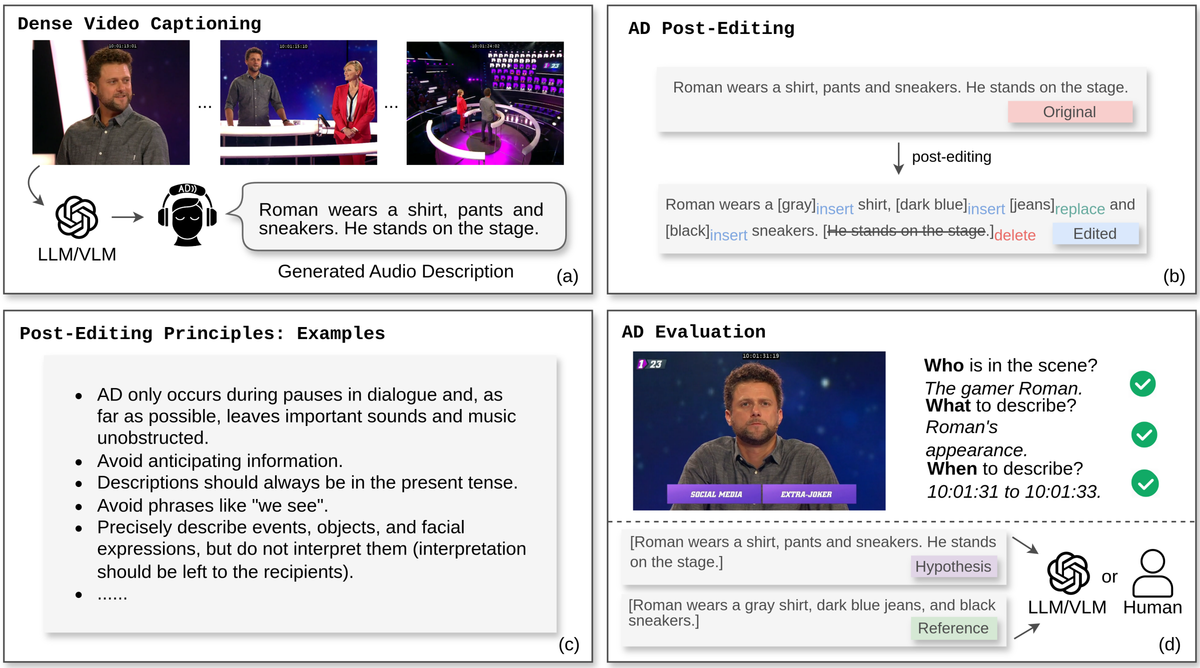


Figure 2: Components of modern AD generation systems with LLM/VLM participation. **(a) Dense Video Captioning**: the task of generating AD scripts from the given video clips; **(b) AD Post-editing**: the task of polishing the generated AD scripts; **(c) AD Script Creation Guidelines**: used as guidance for Post-editing; **(d) AD Evaluation**: Quality assessment of generated ADs. Example taken from the Swiss TV show *1 gegen 100*.

Mazur and Chmiel, 2021; Yan and Luo, 2023). In addition, depending on the provider, AD scripts are sometimes required in multiple languages. Consequently, a shortage of qualified describers exists, leading to high service costs and unmet demand for accessible media.

### 1.2 Motivation

As the demand for AD generation continues to grow due to reinforced legal requirements (Braun and Starr, 2022), both the NLP and CV community have dedicated efforts to solve this problem. In recent years, generative AI technologies such as LLMs (Brown et al., 2020; Touvron et al., 2023) and VLMs (Radford et al., 2021; Ramesh et al., 2022; Li et al., 2023; Zhang et al., 2023a) have demonstrated remarkable capabilities in addressing numerous real-world challenges, including text and image generation. These advancements pave the way for the (semi-)automation of AD generation, as the crucial steps of generating ADs can be offloaded to these large models with significantly less human involvement. Figure 2 depicts the components of a typical modern AD generation system with three crucial steps:

**Dense Video Captioning (DVC)** Given a video, the task is to generate AD scripts that consist of informative descriptions. This inherently multimodal task requires integrating both visual and textual features to create coherent and contextually appropriate AD scripts.

**AD Post-editing (APE)** After generating the initial ADs, refine them according to a set of predefined principles. This post-editing process ensures that the ADs meet specific quality standards and accurately convey the intended information. Note that given sufficient performance of the preceding DVC step, this step would not be necessary; however, the state-of-the-art is such that the APE stage is not yet dispensable.

**AD Evaluation (ADE)** The generated ADs must undergo both quantitative and qualitative assessments, ideally with the involvement of the target groups. This evaluation process measures the effectiveness, accuracy, and overall quality of the ADs, ensuring they meet the necessary criteria for accessibility and usability.

Audio captioning (or audio understanding), a task focused on summarizing or describing auditory information (such as voice effects and environmental sounds), is often misconstrued as part of AD generation. The primary target group for AD generation comprises blind persons and persons with visual impairments, who do not necessarily have hearing impairments. We therefore exclude the discussion of audio captioning.

In this survey, we investigate generative AI technologies for developing AD generation systems, with a **special focus on the participation of LLMs/VLMs**. Specifically, we concentrate on the latest research outcomes in NLP and CV (i.e. papers published from 2020 onward, signifying the release of GPT-3 by OpenAI).

This survey is structured as follows: Section 2 provides a brief review of works on DVC; Section 3 offers an overview of post-editing techniques for AD generation (APE); Section 4 discusses the evaluation of AD generation systems (ADE); Section 5 summarizes the challenges of integrating LLMs/VLMs to real-time AD generation; Section 6 explores future research directions for developing automatic AD generation systems; and Section 7 summarizes the main takeaways of this survey.

## 2 Dense Video Captioning for AD Script Generation

Dense video captioning (DVC) addresses the challenge of establishing connections between clips in videos and their natural language descriptions (Qasim et al., 2023). The term *dense* in DVC signifies the aim to capture as much information as possible to fit the description requirements, which makes DVC a necessary step for AD generation. Typically, DVC outputs multiple sentences as descriptions (Liu and Wan, 2021).

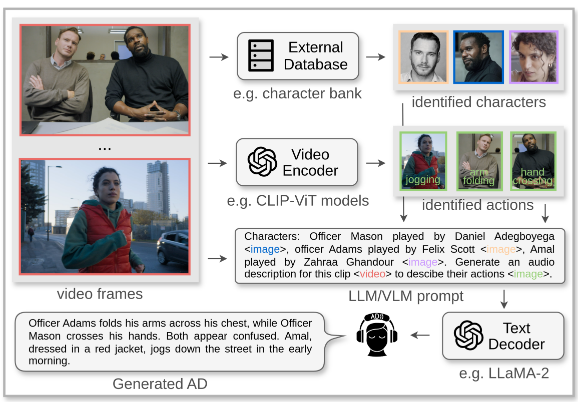


Figure 3: **Process of DVC**: it is often composed of two sub-tasks, i.e., visual feature extraction (VFE)—where a visual encoder decides **whom** and **what** to describe, and dense caption generation (DCG)—where a text decoder works on **how** to describe. Film taken as example: *Baghdad in My Shadow (2019)*.

For the purpose of automatic AD generation, two sub-tasks of DVC are of particular importance:

- **Visual Feature Extraction** (VFE), which involves extracting visual features with a visual encoder within videos that are of interest for DVC. When specialized for AD generation, it means identifying characters (**whom**) and events (**what**) that are important for ADs.
- **Dense Caption Generation** (DCG), which pertains to the methods of automatically generating ADs in the form of natural language scripts derived from the detected event proposals (**how**).

In this survey, we include works on identifying actions, events, and scenes within the context of DVC, as they are all commonly represented in ADs produced by professionals. While actions refer to specific movements (e.g., eating, running, leaving) performed by a subject, typically classified into predefined classes and extracted as bounding boxes within video frames, events can be understood as a series of actions occurring within a temporal range in the video. Scenes, correspondingly, refer to coherent segments of a video that depict a specific event or sequence of actions happening in a continuous time frame, often within a particular setting.

Since ADs are typically inserted during silent moments between dialogues to avoid interference with the ongoing narration—a task that is relatively straightforward—this survey does not delve deeply into techniques for identifying specific video frames for AD insertion. Instead, we focus on reviewing VFE and DCG methodologies to improve AD generation quality, particularly with the integration of LLMs/VLMs.

Next, in Section 2.1 and 2.2, we provide a summary of the relevant VFE and DCG methodologies that can be applied to AD generation. We list relevant studies in Table 1.

### 2.1 Visual Feature Extraction

Convolution-based visual feature extractors (Krizhevsky et al., 2012; Simonyan, 2014; He et al., 2016) were the mainstream of computer vision research for a long time. In recent years, the Vision Transformer (ViT; Dosovitskiy et al. (2020)) has emerged as a central component in modern VFE systems and has been integrated into numerous multimodal VLMs such as CLIP (Radford et al., 2021). Although not being the first work that tries to apply Transformers for CV tasks,

| Research | Venue | Task | Video Encoder | Text Decoder | Method \| Dataset |
|---|---|---|---|---|---|
| Yun and Ro (2024) | CVPR'24 | VFE | Vanilla ViT | not applicable | SHViT |
| Hassani et al. (2023) | CVPR'23 | VFE | Swin Transformer | not applicable | NAT |
| Chen et al. (2023) | ICCV'23 | VFE | Vanilla ViT | not applicable | EVAD |
| Liu et al. (2023) | CVPR'23 | VFE | EfficientViT | not applicable | EfficientViT |
| Zhao et al. (2022) | CVPR'22 | VFE | Vanilla ViT | not applicable | TubeR |
| Liu et al. (2022) | CVPR'22 | VFE | Swin Transformer | not applicable | Video Swin Transformer |
| Yang et al. (2022) | CVPR'22 | VFE | Vanilla ViT | not applicable | Lite Vision Transformer |
| Wu et al. (2022b) | AAAI'22 | VFE | Vanilla VFE | not applicable | Pale Transformer |
| Korbar and Zisserman (2022) | BMVC'22 | VFE | CLIP-ViT-B/32 | not applicable | CLIP-PAD \| CiA |
| Yin et al. (2022) | CVPR'22 | VFE | DeiT | not applicable | A-ViT |
| Wu et al. (2022a) | CVPR'22 | VFE | MViT-v2 | not applicable | MeMViT |
| Brown et al. (2021b) | ICCV'21 | VFE | ResNet50 | not applicable | MuHPC \| VPCD |
| Huang et al. (2021) | arXiv'21 | VFE | Swin Transformer | not applicable | Shuffle Transformer |
| Liu et al. (2021) | ICCV'21 | VFE | Hierarchical ViT | not applicable | Swin Transformer |
| Rao et al. (2021) | NeurIPS'21 | VFE | ViT/DeiT/LV-ViT | not applicable | DynamicViT |
| Wu et al. (2020) | ECCV'20 | VFE | ResNet50-FPN | not applicable | Context-Aware RCNN |
| Huang et al. (2020b) | ECCV'20 | VFE | not applicable | not applicable | MovieNet |
| Kukleva et al. (2020) | CVPR'20 | VFE | NesNeXt-101 | not applicable | LIReC |
| **Research** | **Venue** | **Task** | **Video Encoder** | **Text Decoder** | **Method \| Dataset** |
| Lin et al. (2024) | ECCV'24 | DCG | CLIP-ViT-B/16 | LlaMA-2 | MovieSeq |
| Chu et al. (2024) | arXiv'24 | DCG | GPT-4V | GPT-4 | LLM-AD |
| He et al. (2024) | CVPR'24 | DCG | CLIP-ViT-G/14 + Q-Former | Vicuna-v1.5 | MA-LLM |
| Luo et al. (2024) | arXiv'24 | DCG | TinyLlaVA (SigLIP) | TinyLlaVA (TinyLlama/Phi-2) | Shotluck Holmes |
| Maaz et al. (2024) | arXiv'24 | DCG | CLIP-ViT-L/14 + InternVideo-v2 | Vicuna-v1.5/LlaMA-3/Phi3-Mini | VideoGPT+ \| VCGBench-Diverse |
| Ye et al. (2024) | COLING'24 | DCG | Video-LlaVA-v0 | LlaMA-2 | MMAD |
| Yue et al. (2024) | arXiv'24 | DCG | VideoChat-2/Qwen-VL | GPT-4V | Movie101v2-(zh/en) |
| Zhou et al. (2024) | CVPR'24 | DCG | CLIP-ViT-L/14 | T5-Base | Streaming DVC |
| Blanco-Fernández et al. (2024) | arXiv'24 | DCG | Deformable Transformer | Deformable Transformer | LVC |
| Xie et al. (2024) | arXiv'24 | DCG | VideoLlaMA-(2) | LlaMA-3/Gemma-2 | AutoAD-Zero \| TV-AD |
| Han et al. (2024) | CVPR'24 | DCG | Q-Former | OPT/LlaMA-2 | AutoAD III \| (CMD/HowTo)-AD |
| Yue et al. (2023) | ACL'23 | DCG | Transformer | Transformer | MNScore \| Movie101-zh |
| Jung et al. (2023) | ACL'23 | DCG | LXMERT | EMT + PDVC | KOFCL |
| Han et al. (2023b) | ICCV'23 | DCG | CLIP-ViT-B/32 | GPT-2 | AutoAD II \| MAD-(t-eval/L-char) |
| Han et al. (2023c) | CVPR'23 | DCG | CLIP-ViT-B/32 | GPT-2 | AutoAD \| MAD-v2/AudioVault |
| Shen et al. (2023) | ICCV'23 | DCG | CLIP-ViT-L/14 | Multimodal Transformer | CoCap |
| Yang et al. (2023a) | ACL'23 | DCG | CLIP-ViT-B/16 | Vanilla Transformer | MultiCapClip |
| Lin et al. (2023a) | arXiv'23 | DCG | GPT-4V | GPT-4 | MM-VID |
| Han et al. (2023a) | arXiv'23 | DCG | CLIP-ViT-L/14 + Q-Former | MiniGPT-4/GPT-4 | Shot2Story20K |
| Soldan et al. (2022) | CVPR'22 | DCG | CLIP-ViT-B/32 | not applicable | MAD |
| Zhang et al. (2022) | EMNLP'22 | DCG | CLIP-ViT-B/32 | Vanilla Transformer | MMN/RMN/RNL \| MovieUN |
| Zhu et al. (2022) | COLING'22 | DCG | CNN | T5 | Seg+Cap |
| Deng et al. (2021) | CVPR'21 | DCG | CNN | Vanilla Transformer | SRG |
| Wang et al. (2021) | ICCV'21 | DCG | Deformable Transformer | Vanilla Transformer + LSTM | PDVC |
| Liu and Wan (2021) | ACL'21 | DCG | BMN | BERT + Vanilla Transformer | VPCSum |
| Zhu and Yang (2020) | CVPR'20 | DCG | CNN + Faster R-CNN | Tangled Transformer | ActBERT |
| Lei et al. (2020) | ECCV'20 | DCG | XML | not applicable | XML \| TVR |
| Fang et al. (2020) | EMNLP'20 | DCG | CNN + LSTM | Vanilla Transformer | V2C-Transformer \| V2C |
| Gurari et al. (2020) | ECCV'20 | DCG | not applicable | not applicable | VizWiz-Captions |
| Lin et al. (2020) | EMNLP'20 | DCG | ECO | Vanilla Transformer | SC-SSL |
| Shigeto et al. (2020) | LREC'20 | DCG | ResNet-152 + ResNeXt-101 | GRU | STAIR Actions |
| Huang et al. (2020a) | AACL'20 | DCG | Multimodal Transformer | Vanilla Transformer | ViTT |

Table 1: A collection of studies related to dense video captioning (DVC). We denote works that introduce a new dataset in yellow, works that propose a new method in blue, and works that deliver both in green.

ViT gained its popularity due to its simple design and scalability.

ViT preserves the foundational architecture of the standard Transformer by mapping an image into a sequence of patches, analogous to text tokens in NLP tasks. These patches are then processed to produce linear embeddings, which serve as the inputs to the standard Transformer encoder. In comparison to convolutional kernels, the self-attention mechanism in ViT can be viewed as a soft convolutional inductive bias, while being capable of effectively capturing global dependencies within the input patches (d'Ascoli et al., 2021; Raghu et al., 2021). This enables ViT models to exhibit exceptional feature extraction capabilities, resulting in its outstanding performance across various CV tasks (Chen et al., 2021; Bhojanapalli et al., 2021; Li et al., 2022; Minderer et al., 2022).

Although ViT-based solutions offer significant advantages, they are often constrained by the high complexity associated with exhaustive self-attention computations. To mitigate this challenge, recent research has concentrated on improving efficiency through the development of advanced self-attention computation techniques (Huang et al., 2021; Liu et al., 2021; Yang et al., 2022; Liu et al., 2022; Wu et al., 2022b; Hassani et al., 2023), dynamic feature selection methods (Rao et al., 2021;

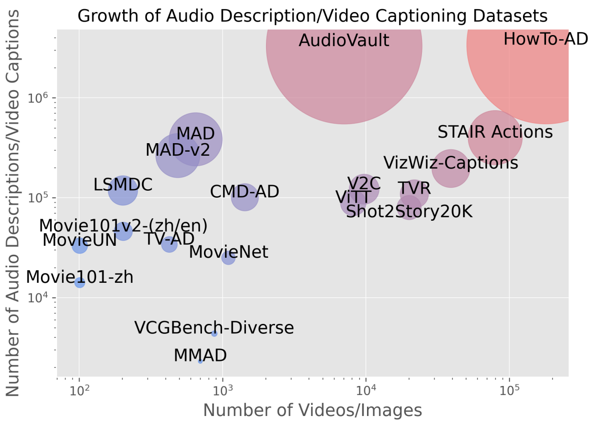


Figure 4: Datasets for AD generation/video captioning. The numbers are visualized in log scale. Red color indicates more recent datasets.

Yin et al., 2022; Chen et al., 2023), and optimized memory scheduling strategies (Wu et al., 2022a; Liu et al., 2023; Yun and Ro, 2024). These approaches are typically evaluated using video action recognition benchmarks, where the system's output is categorized into predefined action classes. AD generation systems often utilize these identified actions as part of the events (**what**) that need to be described.

A crucial additional step for VFE in the context of AD generation is the identification of characters involved in events (Kukleva et al., 2020; Brown et al., 2021b). This process usually involves comparing the extracted features against stored character profiles in an external database (Brown et al., 2021a; Han et al., 2023b), which is essentially an information retrieval task. However, creating and indexing large databases for streaming media content is both costly and often impractical due to copyright concerns. Consequently, by utilizing knowledge encoded in pre-trained VLMs, zero-shot character identification has emerged as a more economical and feasible solution (Bhat and Jain, 2023; Patrício and Neves, 2023; Xie et al., 2024).

### 2.2 Dense Caption Generation

To generate dense video captions from extracted visual features, advanced VLMs are employed to learn the alignment between the generated text tokens and the corresponding visual tokens. For this purpose, multimodal DCG datasets are needed for LLMs to learn the alignment.

Creating large-scale datasets for training is a resource-intensive endeavor. To reduce the workload, researchers often augment existing video datasets with text captions (Lei et al., 2020; Huang et al., 2020a; Gurari et al., 2020; Shigeto et al., 2020; Huang et al., 2020b; Oncescu et al., 2021; Yue et al., 2023; Han et al., 2023a; Yue et al., 2024), or retrieve video counterparts for text annotations (Rohrbach et al., 2017; Soldan et al., 2022; Zhang et al., 2022). Regardless of the annotation approach, subtitles play a crucial role in creating these video-text alignments, often transcribed using automatic speech recognition (ASR) models such as Whisper-based models (Bain et al., 2023; Radford et al., 2023).

In recent years, DCG research has increasingly focused on zero-shot caption generation (Yang et al., 2023a), representing video context as multimodal sequences (Lin et al., 2024), contextualizing visual features using separate image and video encoders (Maaz et al., 2024), developing end-to-end captioning models (Zhu and Yang, 2020; Deng et al., 2021; Wang et al., 2021; Zhu et al., 2022), enhancing model efficiency through memory storage (He et al., 2024; Zhou et al., 2024), and fine-tuning models on well-curated data (Luo et al., 2024). Additionally, efforts have been made to augment ADs with detailed environmental and object information (Fang et al., 2020; Jung et al., 2023; Ye et al., 2024). While these efforts have achieved remarkable performance in generating video captions, they have rarely been fully dedicated to the specific task of AD generation.

Generating high-quality ADs requires the integration of both local context (features within the current video frame) and global context (features from past or future frames). The typical length of movies and other streaming media has caused a trade-off between inference speed, which is particularly critical for live video captioning (Blanco-Fernández et al., 2024), and the quality of the ADs.

To tackle the challenges of curating supervised data and generating high-quality ADs, researchers at the University of Oxford introduced a series of cutting-edge models. In their initial work, Han et al. (2023c) bridge foundation LLM (GPT) and VLM (CLIP) models to perform vision-conditioned AD generation, optimizing the following loss function:

$$\mathcal{L}_{NLL} = -\log p_\Theta \left( \mathcal{T}_{\mathbf{x}_i} | \mathbf{h}_{\mathbf{x}_i}, \mathbf{h}_{\mathrm{AD}}, \mathbf{h}_{\mathrm{Sub}} \right),$$

where the model leverages representations of context frame ($\mathbf{h}_{\mathbf{x}_i}$ from CLIP with $\mathbf{x}_i$ being the current video clip), subtitles ($\mathbf{h}_{\mathrm{Sub}}$), and previous ADs ($\mathbf{h}_{\mathrm{AD}}$) to enhance the generated AD ($\mathcal{T}_{\mathbf{x}_i}$). Thanks to its modular design, the model can be pre-trained even with limited large-scale data for one modality (i.e., visual-only or text-only pre-training). Their AutoAD model demonstrated significant qualitative and quantitative improvements in AD generation.

In their subsequent work (Han et al., 2023b), the authors addressed the character naming issue in AutoAD by introducing a database containing character names, actor profiles, and CLIP face features. Additionally, the authors explored various methods for predicting AD temporal proposals, specifically identifying movie pauses suitable for AD insertion. With these enhancements, their AutoAD-II model achieved further improvements in AD generation quality.

Recently, the authors extended their research with the publication of AutoAD-III (Han et al., 2024), introducing two new AD datasets created from raw videos with soundtracks, a novel Q-Former-based architecture for AD generation, and two new AD evaluation metrics. This work underscores the advancement of LLM/VLM participation in AD generation.

In their latest work (Xie et al., 2024), the authors explore a two-stage zero-shot approach to AD generation. Initially, a VLM is prompted with key information, such as character identities and their interactions, to generate dense captions. These captions are then further summarized into ADs by an LLM. The authors evaluated their AutoAD-Zero model on a custom dataset, TV-AD, achieving competitive results even when compared to supervised models trained on gold-standard ADs.

The AutoAD series of papers illustrate the effectiveness of utilizing LLMs and VLMs for AD generation. Recently, prompt-based pipelines employing GPT-4V as the video encoder and GPT-4 as the text decoder have shown significant potential in producing ADs that align with human production standards (Lin et al., 2023a; Chu et al., 2024). However, further enhancements in generation quality may require the integration of expert knowledge to achieve more coherent and contextually accurate AD narrations.

# 3 AD Post-editing

While a simple video player and a Word editor may be enough for audio describers to edit AD scripts (Minutella, 2022), a variety of specialized professional AD software is available to enhance the quality and efficiency of this process. These tools include options such as CaptioningStar, VDManager (Gagnon et al., 2010), LiveDescribe (Branje and Fels, 2012), YouDescribe, 3Play Media, LiveVoice, Fingertext, Rescribe (Pavel et al., 2020), Frazier, Stellar, and Audible Sight. These platforms offer advanced features tailored specifically for creating, editing, and managing human- or machine-generated ADs, thus providing significant advantages over more general-purpose tools.

Machine-generated ADs often contain grammatical errors and other undesirable elements. To address this issue, text editing models are developed and trained to improve the quality of these texts. These models typically utilize training data that includes human-simplified or corrected texts (Faltings et al., 2021; Kim et al., 2022; Zhang et al., 2023b). Among these, many LLM-based models are fine-tuned with instructions (Raheja et al., 2023, 2024; Shu et al., 2024; Ki and Carpuat, 2024), while others are trained using semi-autoregressive or non-autoregressive decoding techniques (Mallinson et al., 2022; Agrawal and Carpuat, 2022; Zhang et al., 2023b).

Currently, post-editing is still crucial for ensuring adherence to AD production principles (e.g., Figure 2 (c)). However, we contend that future research should focus on the automation of AD generation, thereby eliminating the need for human post-editing.

# 4 AD Evaluation

## 4.1 Automatic Evaluation

Automatic evaluation of ADs typically involves comparing the generated ADs to the gold standards. Classic text generation metrics are employed to assess: 1) textual relevance through N-gram overlaps (e.g., BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and CHRF (Popović, 2015)); or 2) embedding-based semantic similarity between the generated and ground-truth ADs (e.g., MoverScore (Zhao et al., 2019), BERTScore (Zhang et al., 2020), BARTScore (Yuan et al., 2021), (Ref)CLIPScore (Hessel et al., 2021), EMScore (Shi et al., 2022), and (Ref)PAC-S (Sarto et al., 2023)).

Given the multimodal nature of AD generation, image and video captioning metrics are also widely employed to evaluate the quality of generated ADs. Unlike those traditional text generation metrics, CIDEr (Vedantam et al., 2015) assesses N-gram overlaps between generated captions and a set of reference captions, under the assumption that effective machine-generated captions should resemble those produced by a diverse group of humans. SPICE (Anderson et al., 2016) evaluates captions by converting them into graph structures and comparing their semantic propositions. SPIDEr (Liu et al., 2017), a linear combination of SPICE and CIDEr, measures both semantic accuracy and syntactic fluency in generated captions. Fujita et al. (2020) introduced SODA, a metric designed to evaluate machine-generated captions based on their effectiveness in describing the video narrative, with particular emphasis on maintaining temporal order and textual coherence of the captions. BERTHA (Lebron et al., 2022), a BERT-based model trained on human-evaluated captions, is designed to maximize the correlation between automatic evaluation and human judgment. Recently, new image captioning metrics based on multimodal language models have been introduced to enhance scoring explainability (Hu et al., 2023; Chan et al., 2023; Lee et al., 2024), further highlighting the growing role of LLMs and VLMs in evaluating machine-generated AD scripts.

Specialized AD evaluation metrics were also explored. Yue et al. (2023) proposed MNScore, which evaluates the AD quality by accounting for both semantic similarity and character name generation. Han et al. (2024) introduced two additional metrics: CRITIC, a coreference-based approach for measuring character recognition performance, and LLM-AD-Eval, a metric based on LLM-prompting that assesses the overall AD quality.

### 4.2 Human Evaluation

Many human evaluation works focus on how AD end users in different countries perceive ADs in terms of their **usefulness**.

Lopez et al. (2018) explored the usefulness of film and television ADs in the UK. The authors noted that while ADs are useful, there is still room for improvement, particularly in terms of personalization and the integration of sound design techniques (which are proven to be effective in their later work (Lopez et al., 2021)), which could potentially create a more immersive experience for AD end users. Reviers (2018) analyzed Dutch films conducted in Flanders and the Netherlands and confirmed the found of idiosyncratic language patterns. Ferziger et al. (2020) examined the reception of ADs in cultural events in Israel, such as theater patrons, where participants reported high levels of overall satisfaction with the AD services provided. Bausells-Espín (2022) explored the reception of ADs in a pedagogical setting for foreign language (Spanish) teaching. Their study found that, depending on the students' perceived level of difficulty, ADs can be highly helpful in developing transferable and communicative skills such as summarizing and narrating. Arias-Badia and Matamala (2023) found that AD scripts in Catalan adhere to characteristics of "easy-to-understand" language, utilizing simple syntax and lexicon. Yang et al. (2023b) conducted a systematic study on the availability and reception of AD services in mainland China, revealing that, despite significant challenges such as a shortage of AD professionals, limited foundational research, and copyright constraints, AD end users expressed satisfaction with the quality of services even though their comprehension of the movies remained low. Leong et al. (2023) investigated the effectiveness of ADs in aiding blind and visually impaired individuals to interact within 3D virtual environments. The authors concluded that ADs alone are insufficient for facilitating navigation and orientation in such environments and recommended the integration of additional auditory cues such as sound landmarks.

Other relevant studies focus on evaluating the nature of ADs themselves, rather than their functions and effects. For example, Jekat and Carrer (2018) compare the reception of two distinct AD styles: **descriptive** and **interpretative**, among German-speaking AD end users. The study found that, contrary to expectations, users reported a more immersive movie experience with interpretative ADs. This finding challenges the traditional preference for descriptive ADs, which have long been the standard among many German public broadcasters. Gallego (2020) investigated the extent to which subjective ADs in art museums are preferred by blind and visually impaired individuals. Their methodology, which integrates cognitive linguistics and art theory, offers valuable insights into how subjective ADs can effectively enhance guided tours in art museum settings. By contrast, Muñoz (2023) focused on analyzing objectivity in the ADs of Spanish Netflix videos. The results indicate that these ADs are neither purely objective nor entirely subjective. Wang et al. (2022) proposed six distinct methods for assessing the emotional responses of AD users during museum tours.

Human evaluation of ADs often spans multiple research domains, including psychology, pedagogy, and cognitive science, employing methodologies that range from traditional questionnaire-based approaches to measuring neural activities such as EEG signals. However, these interdisciplinary insights have yet to be integrated into the AD generation process using LLM and VLM models. We therefore advocate for multidisciplinary collaborations between AI and non-AI communities to jointly address the challenges in AD generation, reduce technical barriers, and adhere more closely to user-centered principles.

## 5 Challenges of LLM/VLM Integration

### 5.1 Real-time AD Generation

State-of-the-art LLMs and VLMs are practically expensive as they often require either large on-site computing power or stable cloud deployment (Zhang et al., 2024; Qu et al., 2025), making AD generation difficult to achieve real-time performance, especially for high-resolution video streams (Chang et al., 2024) or live events (Di Giovanni et al., 2018; Wilt and Farbood, 2019).

Research in Parameter-efficient Transfer Learning (Houlsby et al., 2019) such as Adapters (Hu et al., 2021; Dettmers et al., 2024) and Prefix-tuning (Li and Liang, 2021) have significantly lower the barrier of fine-tuning LLMs/VLMs under resource-constrained scenarios (Cai et al., 2024). In theaters and sports events, where high responsiveness and precise vision-text alignments are essential, deploying LLMs/VLMs on edge devices such as smartphones remains a significant technical challenge (Qin et al., 2024). These challenges arise primarily due to the limited memory, computational power, and bandwidth of edge devices, as they are difficult to manage the substantial overhead of LLMs/VLMs, even when parameter-efficient techniques are applied (Lin et al., 2023b).

### 5.2 Real-time User Feedback

User feedback is crucial for adjusting the speed, style, level of detail, language, voice, and genre (e.g., movie, sports, lecture, etc.) to maximize user experience. While generating of high-quality ADs has been a prominent focus of research, the development of effective human-computer interaction (HCI) for collecting real-time user feedback remains relatively understudied.

Obtaining meaningful user feedback in real-time from blind persons and persons with visual impairments introduces technical challenges. Traditionally, keyboard-based interaction is a primary mechanism for gathering user input in many scenarios. For example, Natalie et al. (2024) proposed CustomAD, an interface leveraging keyboard navigation to enable users to customize AD generation settings. Similarly, Ning et al. (2024) developed SPICA, an AI-powered system designed to facilitate video exploration using arrow keys for blind persons and persons with visual impairment.

Nevertheless, relying on keyboard interaction is less practical for collecting user feedback on edge devices such as smartphones, where alternative input methods such as voice control may be more suitable and efficient (Szarkowska, 2011; Yamamoto et al., 2024).

AD research related to HCI design must ensure the accessibility of the feedback mechanism itself, addressing diverse user needs and preferences, and managing potential cognitive load during real-time interactions. Addressing these challenges is critical to advancing adaptive AD systems that can enhance user satisfaction and inclusiveness in real-world applications.

## 6 Future Research

AD generation is a complex task that extends far beyond the mere application of LLMs and VLMs. Building on the research reviewed above, we outline the following future research directions.

### 6.1 AD Generation with Human Preferences

Although general international AD standards, such as ISO/IEC TS 20071-21:2015[2], have already existed for a long time, individual nations and audio describers often follow their own inclusive guidelines for AD production (Mazur, 2024). These specific rules have not been incorporated into the tuning process of LLMs/VLMs. Consequently, tuning AD generation systems with these human-crafted guidelines would be beneficial. This could be achieved through LLM alignment techniques (Ouyang et al., 2022; Rafailov et al., 2024; Meng et al., 2024; Ethayarajh et al., 2024), where AD generation models are optimized to produce outputs that align with human preferences.

### 6.2 Personalized AD Generation

Recent research indicates that varying degrees of visual impairment can significantly influence per-

[2] https://www.iso.org/standard/63061.html

ception of ADs (Sève and Horst, 2024), which underscores the importance of personalizing AD generation according to individual requirements of end users (Natalie et al., 2024; Cheema et al., 2024). In addition, AD generation systems for movies and TV episodes should differ from those for art galleries and museums. Moreover, AD generation systems tailored for individuals without intellectual disabilities should be distinct from those intended for persons with intellectual disabilities. Combining AD generation with text simplification could further enhance accessibility for diverse audiences (Braun and Starr, 2021).

Last but not least, AD generation systems should also prioritize scenarios such as higher education, where ADs are crucial in supporting blind students and students with visual impairments for better learning experience.

### 6.3 Machine Translation of AD Scripts

Given that ADs are often available in only one language, research has focused on utilizing machine translation models to translate ADs from one language to another (Matamala and Ortiz-Boix, 2016). This approach aims to facilitate the production of ADs in situations where multilingual audio describers are not available. Fernández-Torné and Matamala (2016) tested machine translation models on English-Catalan AD script pairs, while Vercauteren et al. (2021) conducted similar research with AD script pairs of English-Dutch. Matamala and Villegas (2016) built a multilingual multimodal corpus for ADs. Torné (2016) presented an evaluation of five English-Catalan AD translation systems, employing both automatic and subjective post-editing metrics to assess their performance. These studies not only confirmed the potential of machine translation models for AD translation but also highlighted the significant human post-editing efforts required to achieve satisfactory quality.

The most relevant research in AD translation with LLMs/VLMs is SwissADT, proposed by Fischer et al. (2024), the first multilingual and multimodal AD translation system designed specifically for translating AD scripts in Switzerland's three main languages by utilizing LLMs and incorporating visual inputs from video clips. Their system uses data collected from Swiss national television and synthetic ADs generated with DeepL, demonstrating improved translation quality through both automatic and human evaluations.

However, while promising strides have been made in AD translation research, it remains underexplored and not yet fully integrated into AD production pipelines. More research is needed to refine these models and establish their role in practical applications, ensuring they meet the high standards required for AD production.

## 7 Conclusions

As an inclusive product, ADs have greatly enhanced access to information for blind persons and persons with visual impairments. However, traditional AD production, which relies on human audio describers, is often both costly and time-consuming. In contrast, generative AI technologies, such as LLMs and VLMs, have shown significant potential in automating the AD generation process. In this survey, we reviewed the technologies that are applicable to AD generation, including dense video captioning, (automatic) post-editing, and AD evaluation. As emphasized by Hirvonen et al. (2023), AD production should adhere to user-centric principles, and we believe that LLMs and VLMs can play a crucial role in supporting this requirement.

## Limitations

Our study has two main limitations: 1) We did not explore other DVC sub-tasks, such as video temporal grounding, which involves associating a natural language query with a specific temporal video segment. This omission is because ADs are meant to serve as the final output of AD generation systems, not as queries for retrieving content in videos that blind individuals or individuals with visual impairment cannot perceive. However, we acknowledge that blind individuals and individuals with visual impairment may have information retrieval needs, such as revisiting previous clips in a video, potentially using voice commands. Unfortunately, we found no relevant literature addressing this problem; 2) Given that generated ADs are typically inserted during silent moments between dialogues to avoid interfering with the ongoing narration, this survey does not thoroughly examine techniques for identifying suitable pauses for AD insertion.

## Acknowledgments

This work is supported by the Swiss Innovation Agency Innosuisse, Flagship Inclusive Information and Communication Technology (IICT), funding no. PFFS-21-47. We thank all anonymous reviewers for their constructive feedback.

## References


3Play Media. 3Play Media: Where Innovation Meets Media Accessibility. https://www.3playmedia.com/.

Sweta Agrawal and Marine Carpuat. 2022. An Imitation Learning Curriculum for Text Editing with Non-autoregressive Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7550–7563, Dublin, Ireland. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Blanca Arias-Badia and Anna Matamala. 2023. Audio Description from an Easy-to-understand Language Perspective: A Corpus-based Study in Catalan. *Jostrans: The Journal of Specialised Translation.(nd).*

Audible Sight. Audible Sight: AI-Powered Audio Description Software. https://www.audiblesight.ai/.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-accurate Speech Transcription of Long-form Audio. *INTERSPEECH 2023*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Adriana Bausells-Espín. 2022. Audio Description as a Pedagogical Tool in the Foreign Language Classroom: An Analysis of Student Perceptions of Difficulty, Usefulness and Learning Progress. *Journal of Audiovisual Translation*, 5(2):152–175.

Aaditya Bhat and Shrey Jain. 2023. Face Recognition in the Age of Clip & Billion Image Datasets. *arXiv preprint arXiv:2301.07315*.

Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. 2021. Understanding Robustness of Transformers for Image Classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241.

Eduardo Blanco-Fernández, Carlos Gutiérrez-Álvarez, Nadia Nasri, Saturnino Maldonado-Bascón, and Roberto J López-Sastre. 2024. Live Video Captioning. *arXiv preprint arXiv:2406.14206*.

Carmen J Branje and Deborah I Fels. 2012. LiveDescribe: Can Amateur Describers Create High-Quality Audio Description? *Journal of Visual Impairment & Blindness*, 106(3):154–165.

Sabine Braun and Kim Starr. 2021. *Innovation in Audio Description Research*. Routledge New York.

Sabine Braun and Kim Starr. 2022. Automating Audio Description. In *The Routledge Handbook of Audio Description*, pages 391–406. Routledge.

Andrew Brown, Ernesto Coto, and Andrew Zisserman. 2021a. Automated Video Labelling: Identifying Faces by Corroborative Evidence. In *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 77–83. IEEE.

Andrew Brown, Vicky Kalogeiton, and Andrew Zisserman. 2021b. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3184–3194.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Fenglong Cai, Dong Yuan, Zhe Yang, and Lizhen Cui. 2024. Edge-LLM: A Collaborative Framework for Large Language Model Serving in Edge Computing. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 799–809. IEEE.

CaptioningStar. Audio Description Services by CaptioningStar. https://www.captioningstar.com/audio-description-services/.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646.

Ruei-Che Chang, Yuxuan Liu, and Anhong Guo. 2024. WorldScribe: Towards Context-aware Live Visual Descriptions. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Maryam Cheema, Hasti Seifi, and Pooyan Fazli. 2024. Describe Now: User-driven Audio Description for Blind and Low Vision Individuals. *arXiv preprint arXiv:2411.11835*.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. CrossvViT: Cross-attention Multi-scale Vision Transformer for Image Classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366.

Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. 2023. Efficient Video Action Detection with Token Dropout and Context Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10388–10399.

Peng Chu, Jiang Wang, and Andre Abrantes. 2024. LLM-AD: Large Language Model based Audio Description System. *arXiv preprint arXiv:2405.00983*.

Luque Colmenero, M Olalla, and Silvia Soler Gallego. 2019. Training Audio Describers for Art Museums. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, 18:166–181.

Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. 2021. Sketch, Ground, and Refine: Top-down Dense Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. QLoRA: Efficient Fine-tuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the Limitations of Reference-free Evaluations of Generated Text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elena Di Giovanni et al. 2018. Audio Description for Live Performances and Audience Participation. *The Journal of Specialised Translation*, 29:189–211.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. 2021. ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. In *International conference on machine learning*, pages 2286–2296. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model Alignment as Prospect Theoretic Optimization. In *Proceedings of the 41st International Conference on Machine Learning*.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text Editing by Command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. Association for Computational Linguistics.

Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Video2Commonsense: Generating Commonsense Descriptions to Enrich Video Captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860.

Anna Fernández-Torné and Anna Matamala. 2016. Machine Translation in Audio Description? Comparing Creation, Translation and Post-Editing Efforts. *SKASE Journal of Translation and Interpretation*, 9(1):64–87.

Naomi Ferziger, Yossi Freier Dror, Lirit Gruber, Sara Nahari, Nofar Goren, Nurit Neustadt-Noy, Noomi Katz, and Asnat Bar-Haim Erez. 2020. Audio Description in the Theater: Assessment of Satisfaction and Quality of the Experience among Individuals with Visual Impairment. *British Journal of Visual Impairment*, 38(3):299–311.

Fingertext. A-TotalAccess: Accessibility and Data on TV. https://www.anglatecnic.com/project/a-totalaccess/.

Lukas Fischer, Yingqiang Gao, Alexa Lintner, and Sarah Ebling. 2024. SwissADT: An Audio Description Translation System for Swiss Languages. *arXiv preprint arXiv:2411.14967*.

Frazier. The Text Editor for Audio Describers. https://www.videotovoice.com/.

Louise Fryer. 2016. *An Introduction to Audio Description: A Practical Guide*. Routledge.

Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. SODA: Story Oriented Dense Video Captioning Evaluation Framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 517–531. Springer.

Langis Gagnon, Claude Chapdelaine, David Byrns, Samuel Foucher, Maguelonne Heritier, and Vishwa Gupta. 2010. A Computer-Vision-assisted System for Videodescription Scripting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 41–48. IEEE.

Silvia Soler Gallego. 2020. Defining Subjectivity in Visual Art Audio Description. In *2020 conference of the American Association for Applied Linguistics (AAAL)*. AAAL.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who are Blind. In *Computer Vision–ECCV*

*2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 417–434. Springer.

Mingfei Han, Xiaojun Chang, Heng Wang, and Linjie Yang. 2023a. Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos. *arXiv preprint arXiv:2312.10300*.

Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023b. AutoAD II: The Sequel-Who, When, and What in Movie Audio Description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13645–13655.

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

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD III: The Prequel-Back to the Pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18164–18174.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. 2023. Neighborhood Attention Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6185–6194.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-augmented Large Multimodal Model for Long-term Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Maija Hirvonen, Marika Hakola, and Michael Klade. 2023. Co-translation, Consultancy and Joint Authorship: User-centred Translation and Editing in Collaborative Audio Description. *Journal of Specialised Translation*, (39):26–51.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient Transfer Learning for NLP. In *International conference on machine learning*, pages 2790–2799. PMLR.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3171–3185.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020a. Multimodal Pretraining for Dense Video Captioning. In *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*, pages 470–490.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020b. 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.

Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, and Bin Fu. 2021. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. *arXiv preprint arXiv:2106.03650*.

Anna Jankowska. 2017. Blended Learning in Audio Description Training. *Między Oryginałem a Przekładem*, (38):101–124.

Susanne Johanna Jekat and Luisa Carrer. 2018. A Reception Study of Descriptive vs. Interpretative Audio Description. In *Second Barrier-Free Communication Conference 2018, Geneva, Switzerland, 9-10 November 2018*, pages 54–57. UNIGE Archive Ouverte.

Yeonjoon Jung, Minsoo Kim, Seungtaek Choi, Jihyuk Kim, Minji Seo, and Seung-won Hwang. 2023. Retrieval-augmented Video Encoding for Instructional Captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8554–8568.

Dayeon Ki and Marine Carpuat. 2024. Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273.

Zae Myung Kim, Wanyu Du, Vipul Raheja, Dhruv Kumar, and Dongyeop Kang. 2022. Improving Iterative Text Revision by Learning Where to Edit from Other Revision Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9986–9999.

Bruno Korbar and Andrew Zisserman. 2022. Personalised Clip or: How to Find Your Vacation Videos. *Proceedings of the 33rd. British Machine Vision Conference*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, 25.

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*, pages 9849–9858.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings to Document Distances. In *International conference on machine learning*, pages 957–966. PMLR.

Luis Lebron, Yvette Graham, Kevin McGuinness, Konstantinos Kouramas, and Noel E O'Connor. 2022. BERTHA: Video Captioning Evaluation Via Transfer-learned Human Assessment. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1566–1575.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An Explainable Reference-free Evaluation Metric for Image Captioning Using a Large Multimodal Model. *arXiv preprint arXiv:2406.06004*.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Peter Leong, Melissa Peterson, and Sarah Espinosa. 2023. Audio Description for 3-Dimensional (3D) Virtual Worlds. In *Toward Inclusive Learning Design: Social Justice, Equity, and Community*, pages 127–138. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring Plain Vision Transformer Backbones for Object Detection. In *European conference on computer vision*, pages 280–296. Springer.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, pages 74–81.

Ke Lin, Zhuoxin Gan, and Liwei Wang. 2020. Semi-supervised Learning for Video Captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1096–1106.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023a. MM-VID: Advancing Video Understanding with GPT-4V (ision). *arXiv preprint arXiv:2310.19773*.

Kevin Qinghong Lin, Pengchuan Zhang, Difei Gao, Xide Xia, Joya Chen, Ziteng Gao, Jinheng Xie, Xuhong Xiao, and Mike Zheng Shou. 2024. Learning Video Context as Interleaved Multimodal Sequences. *arXiv preprint arXiv:2407.21757*.

Zheng Lin, Guanqiao Qu, Qiyuan Chen, Xianhao Chen, Zhe Chen, and Kaibin Huang. 2023b. Pushing Large Language Models to the 6G Edge: Vision, Challenges, and Opportunities. *arXiv preprint arXiv:2309.16739*.

Hui Liu and Xiaojun Wan. 2021. Video Paragraph Captioning as a Text Summarization Task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 55–60.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved Image Captioning via Policy Gradient Optimization of SPIDEr. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, and Yixuan Yuan. 2023. EfficientViT: Memory Efficient Vision Transformer with Cascaded Group Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14430.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video Swin Transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

LiveVoice. LiveVoice: Audio Description Services. https://livevoice.io/en/audio-description.

Mariana Lopez, Gavin Kearney, and Krisztián Hofstädter. 2018. Audio Description in the UK: What Works, What Doesn't, and Understanding the Need for Personalising Access. *British journal of visual impairment*, 36(3):274–291.

Mariana Julieta Lopez, Gavin Kearney, and Krisztian Hofstadter. 2021. Enhancing Audio Description: Inclusive Cinematic Experiences through Sound Design. *Journal of Audiovisual Translation*, pages 157–182.

Richard Luo, Austin Peng, Adithya Vasudev, and Rishabh Jain. 2024. Shotluck Holmes: A Family of Efficient Small-scale Large Language Vision Models For Video Captioning and Summarization. *arXiv preprint arXiv:2405.20648*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding. *arXiv preprint arXiv:2406.09418*.

Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. EdiT5: Semi-autoregressive Text Editing with T5 Warm-start. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2126–2138.

Anna Matamala and Pilar Orero. 2007. Designing A Course on Audio Description and Defining the Main Competences of the Future Professional. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 6.

Anna Matamala and Carla Ortiz-Boix. 2016. Accessibility and Multilingualism: An Exploratory Study on the Machine Translation of Audio Descriptions. *Trans. Revista de Traductología*, 20(20):11–24.

Anna Matamala and Marta Villegas. 2016. Building An Audio Description Multilingual Multimodal Corpus: the VIW Project. *MMC2016 (Multimodal Corpora: Computer vision and language processing)*, (11).

Iwona Mazur. 2020. Audio Description: Concepts, Theories and Research Approaches. *The Palgrave handbook of audiovisual translation and media accessibility*, pages 227–247.

Iwona Mazur. 2024. Same Film, Different Audio Descriptions: Describing Foreign Films from a Functional Perspective. *Universal Access in the Information Society*, 23(2):511–521.

Iwona Mazur and Agnieszka Chmiel. 2021. Audio Description Training: A Snapshot of the Current Practices. *The Interpreter and Translator Trainer*, 15(1):51–65.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple Preference Optimization with a Reference-free Reward. *arXiv preprint arXiv:2405.14734*.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple Open-vocabulary Object Detection. In *European Conference on Computer Vision*, pages 728–755. Springer.

Vincenza Minutella. 2022. Audio Description Software Tools. In *The Routledge Handbook of Audio Description*, pages 331–352. Routledge.

Alejandro Romero Muñoz. 2023. Multimodal Analysis as a Way to Operationalise Objectivity in Audio Description: A Corpus-based Study of Spanish Series on Netflix. *Journal of Audiovisual Translation*, 6(2):8–32.

Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, Anhong Guo, and Kotaro Hara. 2024. Audio Description Customization. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–19.

Zheng Ning, Brianna L Wimer, Kaiwen Jiang, Keyi Chen, Jerrick Ban, Yapeng Tian, Yuhang Zhao, and Toby Jia-Jun Li. 2024. SPICA: Interactive Video Content Exploration through Augmented Audio Descriptions for Blind or Low-Vision Viewers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2021. QuerYD: A Video Dataset with High-quality Text and Audio Narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Cristiano Patrício and João C Neves. 2023. Zero-shot Face Recognition: Improving the Discriminability of Visual Face Features Using a Semantic-Guided Attention Model. *Expert Systems with Applications*, 211:118635.

Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 747–759.

Maja Popović. 2015. chrF: Character N-gram F-score for Automatic MT Evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Iqra Qasim, Alexander Horsch, and Dilip K Prasad. 2023. Dense Video Captioning: A Survey of Techniques, Datasets and Evaluation Protocols. *arXiv preprint arXiv:2311.02538*.

Ruiyang Qin, Dancheng Liu, Chenhui Xu, Zheyu Yan, Zhaoxuan Tan, Zhenge Jia, Amir Nassereldine, Jiajie Li, Meng Jiang, Ahmed Abbasi, et al. 2024. Empirical Guidelines for Deploying LLMs onto Resource-constrained Edge Devices. *arXiv preprint arXiv:2406.03777*.

Guanqiao Qu, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang. 2025. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *IEEE Communications Surveys & Tutorials*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-scale Weak Supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, 36.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks? *Advances in neural information processing systems*, 34:12116–12128.

Vipul Raheja, Dimitris Alikaniotis, Vivek Kulkarni, Bashar Alhafni, and Dhruv Kumar. 2024. mEdIT: Multilingual Text Editing via Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 979–1001.

Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. 2023. CoEdIT: Text Editing by Task-specific Instruction Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5274–5291.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-conditional Image Generation with Clip Latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. *Advances in neural information processing systems*, 34:13937–13949.

Nina Reviers. 2016. Audio Description Services in Europe: An Update. *JoSTrans: the journal of specialised translation.-London, 2003, currens*, (26):232–247.

Nina Reviers. 2018. Studying the Language of Dutch Audio Description: An Example of A Corpus-based Analysis. *Translation and Translanguaging in Multilingual Contexts*, 4(1):178–202.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision*, 123:94–120.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6914–6924.

Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. 2023. Accurate and Fast Compressed Video Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. EMScore: Evaluating Video Captioning via Coarse-grained and Fine-grained Embedding Matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17929–17938.

Yutaro Shigeto, Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. 2020. Video Caption Dataset for Describing Human Actions in Japanese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4664–4670.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. RewriteLM: An Instruction-tuned Large Language Model for Text Rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18970–18980.

Karen Simonyan. 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. *International Conference on Learning Representations*.

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*, pages 5026–5035.

Stellar. Stellar 2: AUDIO DESCRIPTION SCRIPTING, EDITING, RECORDING AND MIXING – IN A BROWSER. https://yellaumbrella.tv/stellar-2-home-page/stellar-2-audio-description/.

Agnieszka Szarkowska. 2011. Text-to-speech Audio Description: Towards Wider Availability of AD. *The Journal of Specialised Translation*, 15(1):142–162.

Esther Sève and Dorit Horst. 2024. Studie zu den Bedürfnissen bzgl. nationaler audiovisueller Medienangebote von Menschen mit einer Sehbeeinträchtigung oder Blindheit. In *Endbericht: Studie zu den Bedürfnissen bzgl. audiovisueller Medien*. SZBLIND.

Anna Fernández T Torné. 2016. Machine Translation Evaluation through Post-editing Measures in Audio Description. *InTRAlinea: Online Translation Journal*, 18.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

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.

Gert Vercauteren, Nina Reviers, and Kim Steyaert. 2021. Evaluating the Effectiveness of Machine Translation of Audio Description: the Results of Two Pilot Studies in the English-Dutch Language pair= Avaluació de l'efectivitat de la traducció automàtica de l'àudiodescripció: els resultats de dos estudis pilot per a la combinació lingüística anglès-holandès. *Revista Tradumàtica: Traduccio i Tecnologies de la Informacio i la Comunicacio*, (19):226–252.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end Dense Video Captioning with Parallel Decoding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6847–6857.

Xi Wang, Danny Crookes, Sue-Ann Harding, and David Johnston. 2022. Evaluating Audio Description and Emotional Engagement for BPS Visitors in A Museum Context: An Experimental Perspective. *Translation Spaces*, 11(1):134–156.

Dirk Vander Wilt and Morwaread Mary Farbood. 2019. Automating Audio Description for Live Theater: Using Reference Recordings to Trigger Descriptive Tracks in Real Time. In *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, pages 75–81.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022a. MemViT: Memory-augmented Multiscale Vision Transformer for Efficient Long-term Video Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597.

Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, and Gangshan Wu. 2020. Context-aware RCNN: A Baseline for Action Detection in Videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 440–456. Springer.

Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. 2022b. Pale Transformer: A General vision Transformer Backbone with Pale-shaped Attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2731–2739.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD-Zero: A Training-free Framework for Zero-shot Audio Description. In *Proceedings of the Asian Conference on Computer Vision*, pages 2265–2281.

Ryuichi Yamamoto, Yuma Shirahata, Masaya Kawamura, and Kentaro Tachibana. 2024. Description-based Controllable Text-to-speech with Cross-lingual Voice Control. *arXiv preprint arXiv:2409.17452*.

Jackie Xiu Yan and Kangte Luo. 2023. Audio Description and Interpreting Training: A Comparison of Assessment Criteria from the Perspective of Learners. *Perspectives*, 31(6):1049–1064.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023a. MultiCapCLIP: Auto-encoding Prompts for Zero-shot Multilingual Visual Captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11908–11922.

Chenglin Yang, Yilin Wang, Jianming Zhang, He Zhang, Zijun Wei, Zhe Lin, and Alan Yuille. 2022. Lite Vision Transformer with Enhanced Self-attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11998–12008.

Hui Yang, Weiju Shen, Bo Liu, and Yi Wang. 2023b. A Pilot Study on the Availability and Reception of Audio Description in the Chinese Mainland. *Frontiers in Communication*, 8:1114853.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD: Multi-modal Movie Audio Description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. A-ViT: Adaptive Tokens for Efficient Vision Transformer. In *Proceedings of the IEEE/CVF conference* *on computer vision and pattern recognition*, pages 10809–10818.

YouDescribe. YouDescribe: Audio Description for YouTube Videos. https://youdescribe.org/.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A New Movie Understanding Benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684.

Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. 2024. Movie101v2: Improved Movie Narration Benchmark. *arXiv preprint arXiv:2404.13370*.

Seokju Yun and Youngmin Ro. 2024. SHViT: Single-head Vision Transformer with Memory Efficient Macro Design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An Instruction-tuned Audio-visual Language Model for Video Understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

Mingjin Zhang, Xiaoming Shen, Jiannong Cao, Zeyang Cui, and Shan Jiang. 2024. EdgeShard: Efficient LLM Inference via Collaborative Edge Computing. *IEEE Internet of Things Journal*.

Qi Zhang, Zihao Yue, Anwen Hu, Ziheng Wang, and Qin Jin. 2022. MovieUN: A Dataset for Movie Understanding and Narrating. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1873–1885, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yu Zhang, Yue Zhang, Leyang Cui, and Guohong Fu. 2023b. Non-autoregressive Text Editing with Copy-aware Latent Alignments. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7075–7085.

Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. 2022. Tuber: Tubelet Transformer for Video Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13598–13607.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming Dense Video Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252.

Linchao Zhu and Yi Yang. 2020. ActbBERT: Learning Global-Local Video-text Representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

Wanrong Zhu, Bo Pang, Ashish V. Thapliyal, William Yang Wang, and Radu Soricut. 2022. End-to-end Dense Video Captioning as Sequence Generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5651–5665, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

# A Mathematical Details of Automatic Evaluation Metrics

## A.1 Metrics Based-on N-gram Matching

Text generation metrics can be used to evaluate N-gram overlaps between the ground truth and the generated ADs.

**BLEU** BLEU (Papineni et al., 2002) calculates the precision of unigram, bigram, trigram, and 4-gram matches between the generated text and reference texts.

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

here $p_n$ represents the N-gram precisions, $w_n$ the weight for each N-gram order, and BP the brevity penalty. As one of the most widely used evaluation metric for machine translation, BLEU however sometimes favors shorter, generic ADs, even when longer and more detailed ADs would be more informative for the target group.

**METEOR** METEOR (Banerjee and Lavie, 2005) enhances BLEU by incorporating flexible N-gram matching, including paraphrasing, stemming, and synonym recognition.

$$\mathrm{METEOR} = F_{mean} \cdot (1 - \mathrm{Penalty}),$$

here $F_{mean}$ is a harmonic mean of matching precision and recall (which is usually weighted 9 times more than precision), and $\text{Penalty}$ is a chunk penalty to account for fluency by penalizing scattered word alignments.

**ROUGE-L** ROUGE-L (Lin, 2004) is a recall-focused evaluation metric that measures the longest common subsequence (L) between a generated text and a reference text. Unlike BLEU which relies on N-gram precision, ROUGE-L captures sentence-level fluency and structure by evaluating how well a machine-generated output aligns in order with the reference.

$$\mathrm{Rouge-L} = \frac{(1+\beta^2)P_L R_L}{P_L + \beta^2 R_L},$$

here $\beta$ is a weight factor assigned to precision $P_L$ and recall $R_L$ of the longest common subsequence and is usually set to 1. ROUGE-L is commonly used to evaluate machine-generated text summarizations. Since it prioritizes the longest common subsequences, it may overemphasize longer texts at the expense of conciseness and correctness, in contrast to BLEU.

**CHRF** CHRF (Popović, 2015) computes the F-score of character-level N-gram matches.

$$\mathrm{CHRF} = \frac{(1+\beta^2)PR}{\beta^2 P + R},$$

here $\beta$ is a weight factor usually set to 1. CHRF is more robust to evaluate machine-generated texts of morphologically rich languages. However, CHRF is a surface metric that ignores the contextualized semantics between the candidate and reference text. Valid candidate AD like *"A person wearing a crimson jacket crosses the road."* will receive a low CHRF score compared to the reference *"A man in a red coat walks across the street."*.

**CIDEr** CIDEr (Vedantam et al., 2015) measures how well a generated caption $c_i$ aligns with a set of multiple human-annotated reference captions $S_i = \{s_{i1}, \dots, s_{im}\}$, emphasizing consensus and informativeness.

$$\mathrm{CIDEr}(c_i, S_i) = \sum_{n=1}^{N} w_n \cdot \mathrm{CIDEr}_n(c_i, S_i),$$

where

$$\mathrm{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_{j} \frac{\mathbf{g^n}(c_i) \cdot \mathbf{g^n}(s_{ij})}{||\mathbf{g^n}(c_i)|| \cdot ||\mathbf{g^n}(s_{ij})||},$$

with $\mathbf{g^n}(\cdot)$ being the TF-IDF weightening vector.

**SPICE** Differs from N-gram overlap metrics, SPICE (Anderson et al., 2016) parses both candidate caption $c$ and reference caption $S$ into structured meaning representations (i.e., scene graphs) and count the matching tuples.

$$\mathrm{SPICE}(c, S) = \frac{2 \cdot P(c,S) \cdot R(c,S)}{P(c,S) + R(c,S)},$$

which is essentially F1 measure between precision and recall of matched graph tuples.

### A.2 Metrics Based-on Semantic Matching

A second family of metrics leverages text embeddings to assess the semantic relevance between the candidate and reference texts. These metrics offer the advantage of capturing contextualized meaning, aligning more closely with human judgments, and are therefore commonly used in image and video captioning tasks.

**MoverScore** MoverScore (Zhao et al., 2019) extends the idea of Word Mover's Distance (Kusner et al., 2015) by computing the minimum cost of transforming the candidate text into the reference text using word embeddings. As an optimal transport problem, MoverScore can be formulated as

$$\mathrm{MoverScore}(x^n, y^n) := \min_{F \in \mathbb{R}^{|x^n| \times |y^n|}} \langle C, F \rangle,$$
$$\text{s.t. } F\mathbf{1} = f_{x^n}, F^\top \mathbf{1} = f_{y^n}.$$

where $F$ is the transportation flow matrix denoting the amount of flow transporting from N-grams in candidate text to the reference text. $C$ is the transportation cost matrix with entries being the Euclidean distances between the contextualized word embeddings.

**BERTScore** BERTScore (Zhang et al., 2020) compares contextualized word embeddings between the reference and candidate texts using the pre-trained language model BERT (Devlin et al., 2019) and produce F1 measure from greedily computed precision and recall.

$$R_{\mathrm{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j,$$

$$P_{\mathrm{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j,$$

$$F_{\mathrm{BERT}} = 2 \cdot \frac{P_{\mathrm{BERT}} \cdot R_{\mathrm{BERT}}}{P_{\mathrm{BERT}} + R_{\mathrm{BERT}}}.$$

Here $x$ and $\hat{x}$ represent sequence of reference and candidate tokens. However, using an encoder-only model as the evaluation backbone for decoded texts may introduce application mismatches, as it does not account for fluency, coherence, or decoding-specific artifacts present in generative outputs (Deutsch et al., 2022).

**BARTScore** Unlike BERTScore which relies on embedding-based similarity matching using pre-trained text encoders, BARTScore (Yuan et al., 2021) evaluates text generation based on the likelihood under a pre-trained BART model (Lewis et al., 2020), making it context-aware and fluent-focused.

$$\text{BARTScore} = \sum_{t=1}^{m} \omega_t \log p(\mathbf{y}_t | \mathbf{y}_{<t}, \mathbf{x}, \theta),$$

here the weight $w_t$ can be initialized by inverse document frequencies (IDF).

**BERTHA** Lebron et al. (2022) introduced BERTHA, a BERT-based model equipped with a scoring head trained on human evaluations of machine-generated captions to enhance alignment with human judgment. The model is optimized to maximize the Pearson correlation coefficient with human assessments, ensuring a more accurate and human-aligned evaluation of generated captions.

### A.3 Metrics Based-on Multimodal Alignment

Metrics based on N-gram overlap and semantic relevance are text-only evaluation methods. When applied to the assessment of generated ADs, they fail to capture visual saliency and consistency with the video content, limiting their effectiveness in evaluating the alignment between ADs and visual elements. Image and video captioning metrics are therefore often applied to AD evaluation.

**CLIPScore** CLIPScore (Hessel et al., 2021) is a metric to assess the alignment between texts and images based on the VLM model CLIP (Radford et al., 2021). Given an image embedding $\mathbf{i}$ and a text embedding $\mathbf{t}$ computed by CLIP, the CLIP-Score is defined as their cosine similarity

$$\text{CLIPScore}(\mathbf{t}, \mathbf{i}) = \cos(\mathbf{t}, \mathbf{i}).$$

By including references, CLIPScore can be further extended to include comparison with the references

$$\text{RefCLIPScore}(\mathbf{t}, \mathbf{R}, \mathbf{i}) = \\ \text{H-Mean}(\text{CLIPScore}(\mathbf{t}, \mathbf{i}), \max(\max_{\mathbf{r}\in\mathbf{R}} \cos(\mathbf{t}, \mathbf{r}), 0)),$$

where $\mathbf{R}$ is the set of all reference embeddings, and H-mean is the harmonic mean. CLIPScore can be utilized in both reference-free and reference-based evaluation settings, making it one of the most widely used metrics for image captioning assessment.

**EMScore** Similarly, EMScore (Shi et al., 2022) is a video captioning metric on both coarse- and fine-grained level.

$$\text{EMScore}(X, V) = \\ \frac{1}{2}\left(\text{EMScore}(X, V)_c + \text{EMScore}(X, V)_f\right),$$

where the coarse-grained embedding matching $c$ assesses the overall alignment between the entire video $V$ and the caption $X$, and the fine-grained embedding matching $f$ evaluates the alignment at a more detailed level by comparing individual frames of the video with specific words or phrases in the caption. Vision Transformer and Vanilla Transformer are used as video encoder and text encoder.

**PAC-S** Sarto et al. (2023) introduced a contrastive learning approach called PAC-S to assess the alignment between visual content and generated textual descriptions by incorporating positive-augmented samples during training. PAC-S aims to enhance the evaluation's sensitivity to the nuanced relationship between images or videos and their corresponding captions. Similarly to CLIP-Score, PAC-S can be used in both reference-based (RefPAC-S) and reference-free scenario.

### A.4 Specialized AD Evaluation Metrics

Specialized metrics for AD evaluation have been proposed, often tailored to particular subtasks such as character recognition.

**MNScore** Yue et al. (2023) introduced MNScore, a metric designed to evaluate movie narrations with a particular emphasis on character recognition.

$$\text{MNScore} = \\ \frac{1 \cdot \text{EMScore} + 4 \cdot \text{BERTScore} + 1 \cdot \text{Role-F1}}{6}.$$

By integrating BERTScore and EMScore, MN-Score achieves the highest correlation with human judgment, making it a reliable metric for assessing movie understanding and narration quality.

| Dataset | Language | Use | Pros | Cons |
|---|---|---|---|---|
| AutoAD (Han et al., 2023c) | English | Benchmarking AD generation | Large-scale AD dataset for both training and evaluation | Limited to movie ADs, no fine-grained event segmentation |
| AutoAD-II (Han et al., 2023b) | English | Structured AD evaluation | Improved event alignment and AD coherence compared to AutoAD | Manual annotated character names needed |
| MAD (Han et al., 2023a) | English | Training AD generation models | Dense, time-aligned ADs with high quality | Limited to movies, extensive annotation effort. |
| Movie101 (Yue et al., 2023) | Multilingual | Training and evaluation | Large-scale, rich narrative ADs that cover diverse movie genres | Limited temporal structure, annotation bias affects consistency |
| AutoAD-III (Han et al., 2024) | English | Benchmarking VLM-based AD | One of the largest AD datasets for VLM-based AD generation | Limited to English ADs only |
| TV-AD (Xie et al., 2024) | English | Evaluating TV-series AD | Captures real-world spoken ADs in TV series | Relative small dataset size |
| MMAD (Zhou et al., 2024) | Multilingual | Training multimodal AD models | Covers movies, documentaries, and real-world videos | Requires fine-tuning for AD-specific tasks |
| LLM-AD (Chu et al., 2024) | English | LLM-based AD benchmarking | Evaluates ADs generated by LLMs | May contains LLM-hallucination |

Table 2: Datasets play a critical role in training, fine-tuning, and evaluating AD generation models. We briefly summarize key datasets used in recent AD generation research.

**CRITIC** CRITIC (Han et al., 2024) is designed to evaluate the accuracy of character identification in generated ADs. It employs a co-referencing model to replace ambiguous pronouns (e.g., *he* and *she*) in the ADs with official character names from a pre-defined character bank. The metric then compares the sets of character names in the generated and ground truth ADs, calculating the intersection over union (IoU) to assess accuracy. This approach ensures that the generated ADs correctly identify and reference characters, which is crucial for maintaining narrative coherence in movie descriptions.

**LLM-AD-eval** Proposed in the same work by Han et al. (2024), LLM-AD-eval utilizes LLMs to assess the holistic semantic quality of generated ADs. The evaluation focuses on the alignment between the generated and ground truth ADs concerning human actions, objects, and interactions. LLM-AD-eval scores the generated ADs on a scale from 1 (lowest) to 5 (highest), providing a accurate assessment of their semantic fidelity. This metric leverages the advanced language understanding capabilities of LLMs to evaluate the overall quality and relevance of the ADs.

# B Comparison of Models and Datasets

## B.1 Models for AD Generation

State-of-the-art AD generation models can be categorized into two main classes:

1. End-to-end AD Generation with LLMs/VLMs
   - **Examples**: Video-LLaVA (Ye et al., 2024), GPT-4V (Chu et al., 2024), AutoAD (Han et al., 2023c).
   - **Approach**: These models directly generate ADs by conditioning on visual and textual inputs, leveraging vision-language pre-training.
   - **Advantages**: Ability to handle multimodal understanding by utilizing the zero-shot power of pre-trained LLMs/VLMs.
   - **Limitations**: End-to-end approaches are normally computation-intensive and may lack temporal coherence without explicit feature modeling.
2. Prompt-based LLM for AD Generation
   - **Examples**: AutoAD-Zero (Xie et al., 2024), MovieSeq (Lin et al., 2024).
   - **Approach**: LLMs are prompted with structured visual descriptions to generate coherent ADs.
   - **Advantages**: Text-only fine-tuning making it scalable across different datasets
   - **Limitations**: Lacks fine-grained multimodal grounding which makes the models struggle with visual-text misalignment.

## B.2 Datasets for AD Generation

We summarize major AD generation datasets together with their pros and cons in Table 2.