Bridging Multimodal and Video Summarization: A Unified Survey

Haopeng Zhang
ALOHA Lab, University of Hawaii at Manoa
haopeng.zhang@hawaii.edu

Abstract

Multimodal summarization (MMS) and video summarization (VS) have traditionally evolved in separate communities—natural language processing (NLP) and computer vision (CV), respectively. MMS focuses on generating textual summaries from inputs such as text, images, or audio, while VS emphasizes selecting key visual content. With the recent rise of vision-language models (VLMs), these once-disparate tasks are converging under a unified framework that integrates visual and linguistic understanding. In this survey, we provide a unified perspective that bridges MMS and VS. We formalize the task landscape, review key datasets and evaluation metrics, and categorize major modeling approaches into new taxonomy. In addition, we highlight core challenges and outline future directions toward building general-purpose multimodal summarization systems. By synthesizing insights from both NLP and CV communities, this survey aims to establish a coherent foundation for advancing this rapidly evolving field.

1 Introduction

Summarization is a core task in natural language processing (NLP), traditionally defined as the process of producing a concise and coherent version of a longer source while preserving its essential content (Zhang et al., 2025). With the increasing ubiquity of multimodal content in the digital age, ranging from news articles with images to instructional videos with subtitles, the need to summarize information across multiple modalities has become both urgent and technically feasible. This has given rise to two closely related but historically distinct areas of research: multimodal summarization (MMS) and video summarization (VS).

Multimodal summarization focuses on generating summaries from inputs that combine text, images, audio, and other non-linguistic modalities (Li et al., 2017). It is often approached as a text generation task and studied within the NLP community (Atharva et al., 2023). In contrast, video summarization, which is traditionally rooted in the computer vision (CV) domain, aims to produce a condensed version of a video by selecting keyframes or segments, typically without involving language generation (Apostolidis et al., 2021a). As a result, these two lines of work have evolved in parallel, with limited interaction between their research communities, benchmarks, and methodologies.

However, the growing prevalence of multimodal digital content has catalyzed a convergence between MMS and VS. A pivotal shift in MMS was the introduction of Multimodal Summarization with Multimodal Output (MSMO), a paradigm that generates summaries comprising both text and visuals for a richer digest of information (Zhu et al., 2018b). Concurrently, video summarization has evolved beyond traditional extractive video-to-video (V2V) techniques (Zhang et al., 2016; Ji et al., 2019). It now includes abstractive video-to-text (V2T) generation and hybrid video-to-text+video (V2VT) systems (Lin et al., 2023; Hua et al., 2024). This progression demonstrates a clear alignment with the multimodal output objectives pioneered by MSMO, as both aim to integrate textual narratives with key visual highlights.

Recent breakthroughs in Vision-Language Models (VLMs) such as BLIP-2 (Li et al., 2023b), Flamingo (Alayrac et al., 2022), and GPT-4V are now unifying these once-separate research threads. By capably processing and generating content across text, image, and video modalities, these models enable the creation of truly integrated summarization systems (Argaw et al., 2024). This development mirrors the recent history of text-only summarization, where Large Language Models (LLMs) catalyzed transformative progress (Liu and Lapata, 2019; Zhang et al., 2023b,a). The success of LLMs provides a compelling blueprint for leverag-

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 157–171
November 8, 2025 ©2025 Association for Computational Linguistics

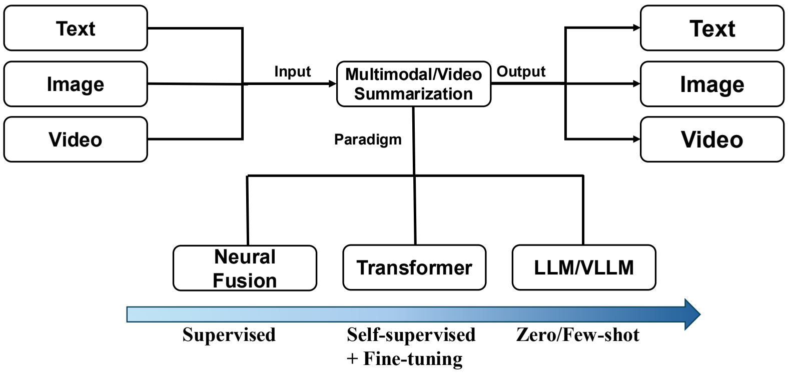

Figure 1: Overview of multimodal/video summarization input/output modalities, and major paradigms shift.

ing VLMs to achieve similar breakthroughs in the multimodal domain.

Despite this convergence, the literature lacks a comprehensive and unified survey that brings together the full landscape of multimodal and video summarization. Existing surveys typically focus either on multimodal summarization from a language perspective (Jangra et al., 2023; Atharva et al., 2023) or on visual summarization from a frame-selection or temporal segmentation perspective (Apostolidis et al., 2021a). Thus, there is a pressing need to synthesize these research directions under a holistic perspective in the LLM era.

This survey bridges this critical gap by providing a comprehensive and unified overview of MMS and VS, focusing on their convergence. We synthesize research from both NLP and CV, examining common challenges, modeling paradigms, datasets, and evaluation metrics under a cohesive lens, as shown in Figure 1. Specifically, this survey:

1) Formalizes the unified task landscape and discusses pertinent evaluation metrics (Section 2);
2) Reviews key datasets and benchmarks from both traditions, grouping by input and output modalities (Section 3);
3) Proposes a new unified taxonomy to categorize existing modeling approaches for MMS and video summarization, highlighting how recent pretrained VLMs are reshaping the summarization landscape (Section 4);
4) Discusses open challenges and charts promising future directions for a truly unified multimodal summarization framework that can reason over and summarize complex, cross-modal content (Section 5).

By unifying these perspectives, we hope to support a broader understanding of multimodal summarization and foster collaboration between NLP and CV researchers working at the intersection of language, vision, and video understanding.

Scope of the Survey This survey provides a comprehensive overview of the evolving MMS and VS fields. We focus specifically on inputs and outputs derived from text, image, and video modalities, excluding others such as audio and tabular data. Throughout the survey, we make a clear distinction: the image modality refers to static, standalone images, while dynamic sequences of frames or clips are treated as video modalities.

2 Background

2.1 Task Definition

At its core, summarization aims to distill salient information from a complex data source $\mathcal{X}$ into a compact representation $\mathcal{Y}$. In the context of this survey, the input source $\mathcal{X}$ is a set of aligned data streams from various modalities, which we formally denote as:

$$\mathcal{X} = \{x^{(T)}, x^{(I)}, x^{(V)}, \dots\}, \qquad (1)$$

where $x^{(T)}$ represents textual content (e.g., transcripts, captions), $x^{(I)}$ represents static images (e.g., photos, diagrams), and $x^{(V)}$ represents video data (e.g., frames, segments).

Historically, multimodal and video summarization have addressed different instances of this general problem. Traditional MMS primarily focused on generating a textual summary $\mathcal{Y} = y^{(T)}$ from a combination of non-video inputs (e.g., $\mathcal{X} = \{x^{(T)}, x^{(I)}\}$). In contrast, traditional VS operated on a single modality ($\mathcal{X} = \{x^{(V)}\}$) to produce a condensed video output $\mathcal{Y} = y^{(V)}$ by selecting keyframes or segments.

The process of generating the summary $\mathcal{Y}$ from the input $\mathcal{X}$ falls into two main categories:

Formulation	Input($\mathcal{X}$)	Output ($\mathcal{Y}$)	Type	Example
T + I → T	$\{x^{(T)}, x^{(I)}\}$	$y^{(T)}$	Abstractive	Summarize news article with photos
T + I → T + I	$\{x^{(T)}, x^{(I)}\}$	$\{y^{(T)}, y^{(I)}\}$	Hybrid	Summarize social media post
V → V	$x^{(V)}$	$y^{(V)}$	Extractive	Video keyframe selection
V + T → T	$\{x^{(V)}, x^{(T)}\}$	$y^{(T)}$	Abstractive	Summarize video with transcript
V + T → T + I	$\{x^{(V)}, x^{(T)}\}$	$\{y^{(T)}, y^{(I)}\}$	Hybrid	Summarize video and select cover frame

Table 1: Common task settings in MMS and VS, categorized by input/output modalities and summarization type.

Modality	Metric	Ref
Textual	ROUGE (Lin, 2004)	✓
	METEOR (Banerjee and Lavie, 2005)	✓
	SummaQA (Scialom et al., 2019)	✗
	BLANC (Vasilyev et al., 2020)	✗
	SUPERT (Gao et al., 2020)	✗
	BERTScore (Zhang et al., 2020)	✓
	GPTScore (Fu et al., 2023)	✗
	G-Eval (Liu et al., 2023)	✗
Visual	Object Overlap (Lee et al., 2012)	✓
	Frame Precision/Recall (Gong et al., 2014)	✓
	Semantic Text Comparison (Yeung et al., 2014)	✓
	Kendall's τ, Spearman's ρ	✓
Multimodal	CLIPScore (Hessel et al., 2021)	✗
	CLIPBERTScore (Wan and Bansal, 2022)	✓
	VT-CLIPScore (Lin et al., 2023)	✗
	FALLACIOUS (Zhang et al., 2024)	both
	mLLM-EVAL (Zhuang et al., 2024)	✗

Table 2: Evaluation Metrics for Multimodal Summarization. 'Ref' indicates whether a ground-truth reference summary is required.

Extractive Summarization. This approach selects a subset of the original input, $\mathcal{Y} = \mathcal{X}' \subset \mathcal{X}$. The goal is to identify and present the most representative components of the source, such as key sentences or video clips.

Abstractive Summarization. This approach generates entirely new content, $\mathcal{Y} = f(\mathcal{X})$, by transforming and synthesizing information from the source. The output is often a novel textual narrative that is not restricted to phrases from the input. It is also possible to combine extractive and abstractive methods to build **hybrid summarization** methods.

This formalization allows us to systematically categorize the full spectrum of MMS and VS tasks based on their input and output modalities, as we detail in Table 1.

2.2 Evaluation Metrics

Evaluating multimodal summaries presents a significant challenge due to the diversity of output formats (text, images, video) and the multifaceted nature of summary quality. Beyond simple relevance, a successful summary must exhibit coherence, fluency, factuality, and, crucially, cross-modal consistency.Table 2 provides an overview of existing evaluation metrics across different modalities.

Given these challenges, most automated evaluation metrics are reference-based. They operate by quantifying the similarity between a machine-generated summary $\mathcal{Y}$ and a set of human-created reference summaries $\mathcal{Y}^*$, generally expressed via a scoring function or similarity measurement $\mathcal{SIM}$:

$$\text{score} = \mathcal{SIM}(\mathcal{Y}, \mathcal{Y}^*). \tag{2}$$

Textual Evaluation For assessing generated text, standard metrics from text summarization are widely adopted. These are often reference-based, beginning with classic methods like **ROUGE** (Lin, 2004) that measures lexical overlap between the generated and reference summaries using n-grams (e.g., ROUGE-N) and longest common subsequences (e.g., ROUGE-L). **METEOR** (Banerjee and Lavie, 2005) is an F-measure based on unigram matching with synonymy. To capture deeper meaning, semantic similarity metrics such as **BERTScore** (Zhang et al., 2020) leverage contextual embeddings from BERT models.

Reference-free alternatives also exist, such as **SummaQA** (Scialom et al., 2019) that provides QA-based factuality assessment, **BLANC** (Vasilyev et al., 2020) that measures utility via masked token prediction, and **SUPERT** (Gao et al., 2020) that uses unsupervised sentence representations. More recently, **LLM-as-a-judge** metrics like **GPTScore** (Fu et al., 2023) and **G-Eval** (Liu et al., 2023) have emerged to employ LLMs to generate aspect-guided or interpretable quality assessments.

However, these metrics often suffer from inability to capture cross-modal grounding. While semantic and LLM-based approaches successfully mitigate the overemphasis on lexical overlap found in older metrics, they remain unable to reason beyond the text modality.

Visual Evaluation Evaluation for visual components, particularly in extractive video summarization (V2V), focuses on the quality of content selection. The standard approach quantifies the overlap between system-selected keyframes or segments and a human-annotated ground truth, typically using **precision, recall, and F1-score** (Gong et al., 2014; Otani et al., 2019). An alternative paradigm, introduced by Otani et al. (2019), evaluates the model's ability to predict frame-level importance scores directly. Instead of measuring set overlap, this method measures the correlation between the model's importance rankings and human judgments using statistical coefficients like **Kendall's** τ (Kendall, 1945) and **Spearman's** ρ (Zwillinger and Kokoska, 1999) (Narasimhan et al., 2021; Saquil et al., 2021). While less common today, some early work also explored pixel-level (Khosla et al., 2013) or object-based (Lee et al., 2012) similarity.

Multimodal Evaluation Evaluating the interplay between modalities is the central challenge of multimodal summarization. Effective metrics must assess not just the quality of individual components but, more importantly, the cross-modal consistency, alignment, and coherence between them. Several approaches have been proposed to address this:

CLIPScore (Hessel et al., 2021) is a foundational metric that leverages the cross-modal embeddings from CLIP (Radford et al., 2021) to quantify the semantic relevance between an image (or video frame) and a textual description, serving as a powerful tool for reference-free grounding.

CLIPBERTScore (Wan and Bansal, 2022) creates a composite score by combining CLIPScore (for image-text relevance) with BERTScore (for text quality), providing a more holistic assessment.

VT-CLIPScore (Lin et al., 2023) adapts CLIPScore specifically for the video-text domain, often with task-specific fine-tuning.

FALLACIOUS (Zhang et al., 2024) introduces metrics for detecting factual inconsistencies in generated text with respect to its corresponding images/videos, offering both reference-based and reference-free variants.

mLLM-EVAL (Zhuang et al., 2024) proposes using multimodal LLMs as evaluators, a method that shows a high correlation with human judgments.

Despite these advances, automated metrics often provide coarse-grained scores and may miss subtle factual inconsistencies or nuanced cross-modal relationships (Hanna and Bojar, 2021). Consequently, **Human Evaluation** remains the gold standard, indispensable for judging high-level qualities like coherence, informativeness, and the overall faithfulness of the summary to the multimodal source.

3 Datasets

The advancement of multimodal summarization has been significantly propelled by the development of diverse and specialized datasets. These resources are crucial for training robust models, benchmarking research progress, and exploring the nuances of different summarization objectives. In this section, we survey key datasets and categorize them by modalities into four groups: (1) Text + Image to Text datasets (TI2T), (2) Video to Video datasets (V2V), (3) Video to Text datasets (V2T), and (4) Multimodal Summarization with Multimodal Output (MSMO) datasets, which involve richer modality combinations. Table 3 provides a detailed comparative overview of these datasets.

3.1 Text + Image to Text Datasets

Early multimodal summarization research focused on augmenting textual documents with associated images, often sourced from news articles, Wikipedia, and e-commerce, where visual and textual information are naturally aligned. Early work includes the **MMSS** dataset (Li et al., 2018), which provides (sentence, image, headline) triples from news sources for generating headline-style summaries. In the e-commerce domain, the large-scale Chinese dataset **EC-product** (Li et al., 2020a) offers product images and descriptions paired with human-written summaries of key features. More recent efforts leverage web-scale data, such as **Wiki-Web2M** (Burns et al., 2023), which sources millions of multimodal sections from Wikipedia and uses the initial sentences as noisy proxy summaries. To improve upon this, **REFINESUMM** (Patil et al., 2024) provides a cleaner version by using multimodal LLMs to generate and filter higher-quality reference summaries from the same source.

3.2 Video to Text Datasets

V2T datasets have evolved significantly, progressing from short-form video captioning to enabling

Dataset	Domain	Size	Language	Input	Output
MMSS (Li et al., 2018)	News	66,000	English	T, I	T
EC-product (Li et al., 2020a)	Product	1,375,453	Chinese	T, I	T
WikiWeb2M (Burns et al., 2023)	Instructional	2,000,000+	English	T, I	T
REFINESUMM (Patil et al., 2024)	Instructional	77,021	English	T, I	T
MSVD (Chen and Dolan, 2011)	Open	1,970	English	V	T
YouCook (Das et al., 2013)	Cooking	88	English	V	T
MSR-VTT (Xu et al., 2016)	Open	7,180	English	V	T
ActivityNetCap (Krishna et al., 2017)	Activities	20,000	English	V	V,T
How2 (Sanabria et al., 2019)	Instructional	80,000	Portuguese + English	V, T	T
VT-SSum (Lv et al., 2021)	Open	1,000	English	V, T	T
StreamHover (Cho et al., 2021)	Stream	500 h	English	V, T	T
Shot2Story20K (Han et al., 2023)	Open	20,000	English	V, T	T
VISTA (Liu et al., 2025)	Academic	18,599	English	V, T	T
SumMe (Gygli et al., 2014)	Events	25	English	V	V
TVSum (Song et al., 2015)	Web video	50	English	V	V
OVP (De Avila et al., 2011)	Documentary	50	English	V	V
VSUMM (De Avila et al., 2011)	Web video	50	English	V	V
LoL (Fu et al., 2019)	E-sports	218	English	V	V
EDUVSUM (Ghauri et al., 2020)	Lectures	98	English	V	V
Ads-1K (Tang et al., 2022)	Commercials	1,041	English	V	V
LfVS-T (Argaw et al., 2024)	Open	1,041	English	V	V
MSMO (Zhu et al., 2018b)	News	314,581	English	T, I	T, I
VMSMO (Li et al., 2020d)	News	184,920	Chinese	V, T	T, I
MM-AVS (Fu et al., 2021)	News	2,173	English	V, T, I	T, I
XMSMO-News (Tang et al., 2023)	News	4,891	English	V, T	T, I
MLASK (Krubiński and Pecina, 2023)	News	41,243	Czech	V, T	T, I
MMSum (Qiu et al., 2024)	Open	5,100	English	V, T	V, T
VideoXum (Lin et al., 2023)	Activities	14,001	English	V	V, T
Instruct-V2Xum (Hua et al., 2024)	Open	30,000	English	V	V, T

Table 3: Summary of existing multimodal and video summarization datasets. The "Input" and "Output" columns use V (Video), T (Text), and I (Image) to indicate modalities.

long-form, abstractive summarization.

Early benchmarks focused on generating single-sentence descriptions for short video clips. These include **MSVD** (Chen and Dolan, 2011) with its open-domain content, the cooking-focused **YouCook** (Das et al., 2013), and the large-scale **MSR-VTT** (Xu et al., 2016). A key step towards more detailed understanding was made with **ActivityNetCap** (Krishna et al., 2017), which provides thousands of untrimmed videos annotated with temporally localized captions for multiple events within each video.

More recent datasets provide richer annotations to support true summarization in complex, real-world domains. **How2** (Sanabria et al., 2019) is a massive corpus of 80,000 instructional videos with aligned transcripts and summaries. For highly granular analysis, **Shot2Story20K** (Han et al., 2023) offers 20,000 clips with shot-level captions, video-level summaries, and optional transcripts. The **VT-SSum** dataset provides around 1,000 video-transcript pairs, with each pair containing manually annotated segment boundaries and corresponding summaries.

Other datasets focus on specific genres, such as **StreamHover** (Cho et al., 2021), which consists of long-form livestream videos from gaming platforms, paired with annotations for highlight detection and key moment spotting, and **VISTA** (Liu et al., 2025) that provides AI conference video presentations paired with their author-written abstracts, enabling textual summarization of academic talks from top AI/ML conferences.

3.3 Video to video Datasets

V2V datasets are foundational for extractive summarization, providing video inputs and human annotations as either frame-level importance scores or complete reference summaries.

The most common benchmarks include **SumMe** (Gygli et al., 2014), a classic dataset of 25 diverse videos with dense importance scores from 15–18 annotators, and **TVSum** (Song et al., 2015), which offers 50 videos across 10 categories (e.g., news, documentaries, vlogs) similarly annotated with importance scores from 20 annotators. In contrast, earlier datasets like OVP and VSUMM (De Avila et al., 2011) each contain 50 short consumer videos annotated with five human-generated summaries.

In addition to these general benchmarks, several datasets target specific domains. For e-sports, **LoL** (League of Legends Highlights) (Fu et al., 2019) provides a large-scale collection of 'League of Legends' gameplay videos paired with professionally edited highlight clips as ground truth. For education, **EDUVSUM** (Ghauri et al., 2020) targets lecture video summarization by providing educational videos and corresponding human-generated key segment summaries. For advertising, **Ads-1K** (Tang et al., 2022) is a large-scale benchmark, containing 1,000 advertisement videos across 10 categories, annotated for highlight detection.

3.4 MSMO Datasets

Datasets for Multimodal Summarization with Multimodal Output (MSMO) enable models to generate summaries that combine both textual and visual (keyframes or clips) elements.

The development of MSMO datasets was initially driven by the NLP community, with a primary focus on the news domain. The foundational **MSMO** dataset (Zhu et al., 2018b) pioneered this by annotating news articles from the CNN/DailyMail corpus with both textual summaries and salient representative images. Building on this, **MM-AVS** (Fu et al., 2021) enriches articles with a wider array of inputs (text, images, video, audio, transcripts) for generating a text summary alongside associated media clips. Similar datasets exist for other languages, including **VMSMO** (Li et al., 2020d), a large-scale Chinese news dataset for generating summaries and selecting cover images, and **MLASK** (Krubiński and Pecina, 2023), a Czech news corpus with human-written summaries and manually selected representative images.

Meanwhile, the VS community has increasingly integrated textual components, creating datasets where video is the primary input. For example, **VideoXum** (Lin et al., 2023) enhances the ActivityNet dataset by linking dense captions to specific keyframes, enabling joint visual-textual summary generation. **XMSMO-News** (Tang et al., 2023) uses BBC News videos as input, with their titles serving as the reference textual summary, requiring the joint selection of representative video content. Most recently, **Instruct-V2Xum** (Hua et al., 2024) provides a large-scale corpus of 30,000 video samples with high-quality annotations specifically designed for instruction-tuned Video-to-Text+Video (V2VT) summarization.

4 Methodology

This section traces the evolution of multimodal and video summarization methods and models across three paradigms. We begin with (1) Early neural architectures that addressed MMS and VS with separate, task-specific models We then examine (2) Transformer-based methods, whose attention mechanisms enabled deeper cross-modal reasoning and integration. Finally, we explore (3) Large Language and Vision-Language Models (LLMs and VLMs), which unify summarization under a single, generative framework. Our analysis focuses on key architectural shifts and the progressive integration of modalities that drove the convergence of these fields.

4.1 Early Neural Architectures

Early research in both MMS and VS was dominated by specialized neural architectures, which typically combined Convolutional Neural Networks (CNNs) for visual feature extraction with Recurrent Neural Networks (RNNs) for processing sequential information.

In the domain of MMS, early systems primarily targeted text-image pairs, employing architectures that paired CNN-based image encoders with RNN-based text decoders. Foundational sequence-to-sequence frameworks by Zhu et al. (2018a), Chen and Zhuge (2018), and Li et al. (2018) highlighted the central challenge: effective cross-modal fusion. To this end, researchers developed sophisticated mechanisms to align visual and textual information, including joint multimodal attention (Zhu et al., 2020a), selective gating to filter irrelevant visual content (Li et al., 2020b), hierarchical correlation modeling (Zhang et al., 2022c), and knowledge distillation to transfer cross-modal insights to more compact models (Zhang et al., 2022d).

In parallel, early VS methods focused primarily on extractive summarization, using RNNs to model temporal dependencies for selecting keyframes or video segments. Prominent examples include the vsLSTM model (Zhang et al., 2016) that combined LSTMs (Graves and Graves, 2012) with Determi-

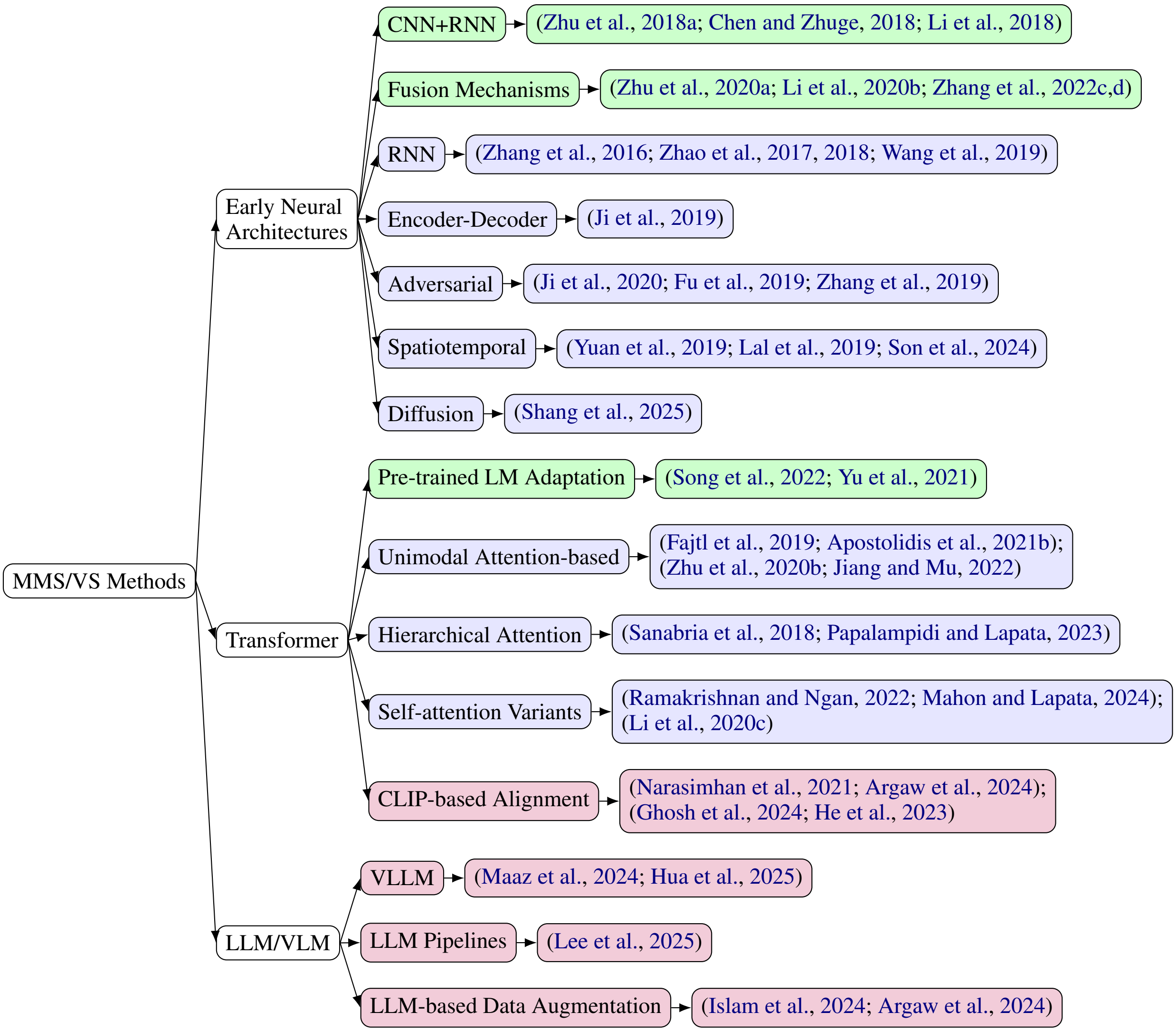
MMS/VS Methods
Early Neural Architectures
CNN+RNN
(Zhu et al., 2018a; Chen and Zhuge, 2018; Li et al., 2018)
Fusion Mechanisms
(Zhu et al., 2020a; Li et al., 2020b; Zhang et al., 2022c,d)
RNN
(Zhang et al., 2016; Zhao et al., 2017, 2018; Wang et al., 2019)
Encoder-Decoder
(Ji et al., 2019)
Adversarial
(Ji et al., 2020; Fu et al., 2019; Zhang et al., 2019)
Spatiotemporal
(Yuan et al., 2019; Lal et al., 2019; Son et al., 2024)
Diffusion
(Shang et al., 2025)
Transformer
Pre-trained LM Adaptation
(Song et al., 2022; Yu et al., 2021)
Unimodal Attention-based
(Fajtl et al., 2019; Apostolidis et al., 2021b);
(Zhu et al., 2020b; Jiang and Mu, 2022)
Hierarchical Attention
(Sanabria et al., 2018; Papalampidi and Lapata, 2023)
Self-attention Variants
(Ramakrishnan and Ngan, 2022; Mahon and Lapata, 2024);
(Li et al., 2020c)
CLIP-based Alignment
(Narasimhan et al., 2021; Argaw et al., 2024);
(Ghosh et al., 2024; He et al., 2023)
LLM/VLM
VLLM
(Maaz et al., 2024; Hua et al., 2025)
LLM Pipelines
(Lee et al., 2025)
LLM-based Data Augmentation
(Islam et al., 2024; Argaw et al., 2024)

Figure 2: Unified taxonomy of MMS and VS Methods across three paradigms: green means MMS, blue means VS, and red means both.

nantal Point Processes (DPP) for diversity, hierarchical RNNs (H-RNN (Zhao et al., 2017), HSA-RNN (Zhao et al., 2018)), stacked LSTMs (Wang et al., 2019), and various encoder-decoder structures (Ji et al., 2019). Another line of work utilized adversarial training to better align generated summaries with ground-truth distributions, as seen in DASP (Ji et al., 2020) and various GAN-based models (Fu et al., 2019; Zhang et al., 2019). Further advancements came from spatiotemporal models like CRSum (Yuan et al., 2019) and MerryGoRoundNet (Lal et al., 2019), which integrated both local and global video cues. Recent research also explores diffusion models (Shang et al., 2025). These approaches were predominantly extractive, aiming to select keyframes or segments. This era was characterized by task-specific architectures where explicit modality fusion and salience modeling remained central research challenges.

4.2 Transformer-based Methods

The advent of the Transformer architecture, along with large-scale pre-trained models like BERT (Devlin et al., 2019) and CLIP (Radford et al., 2021), instigated a paradigm shift. By replacing RNNs with self-attention mechanisms, these models enabled a deeper semantic understanding.

In MMS, the focus shifted toward leveraging large pre-trained language models as powerful backbones. Early approaches adapted text-only models like BART (Lewis et al., 2020) by either converting images into textual descriptions to fit the model's expected input (Song et al., 2022) or by developing methods to directly inject visual features into the language model's architecture (Yu et al., 2021).

In parallel, VS adopted Transformers to overcome the limitations of RNNs in modeling long-range temporal dependencies. Models like VASNet (Fajtl et al., 2019) and PGL-SUM (Apostolidis et al., 2021b) introduced Transformer-based architectures for frame importance scoring with soft self-attention and positional encodings, while others integrated additional modules like salient region detection in DSNet (Zhu et al., 2020b), collaborative learning for moment localization in iPTNet (Jiang and Mu, 2022) and novel spatiotemporal representations for the attention mechanism (Son et al., 2024). This foundation rapidly evolved towards abstractive, video-to-text (V2T) generation. Sophisticated Transformer-based systems, building on earlier hierarchical attention models (Sanabria et al., 2018; Papalampidi and Lapata, 2023), became capable of generating coherent, long-form narratives from video, often through hybrid extractive-abstractive frameworks (Ramakrishnan and Ngan, 2022; Li et al., 2020c; Mahon and Lapata, 2024).

The convergence of MMS and VS was most significantly catalyzed by vision-language models like CLIP. Its shared embedding space for images and text provided a powerful foundation for cross-modal alignment. This breakthrough enabled a new class of models that used CLIP embeddings to ground textual summaries in visual content, seen in methods for both video summarization (CLIP-It (Narasimhan et al., 2021), LfVS-T (Argaw et al., 2024)) and multimodal QA summarization (CLIP-syntel (Ghosh et al., 2024)). Frameworks like A2Summ (He et al., 2023) exemplified this trend by unifying video-text summarization under a single alignment-guided attention module.

By this stage, MMS and VS shared attention-based fusion pipelines, common pre-trained encoders, and were increasingly framed under a unified cross-modal objective

4.3 LLM-driven Multimodal Reasoning

The most recent paradigm shift is driven by the integration of LLMs (Achiam et al., 2023; Touvron et al., 2023) as central reasoning engines, which excel at contextual comprehension (Brown et al., 2020), cross-domain reasoning (Wei et al., 2022; Kojima et al., 2022). The cross modality understanding of LLMs were enabled by Vision-Language Models (VLMs) like Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b). These models established a now-standard architecture: visual features from a dedicated encoder are projected into the LLM's word embedding space, allowing it to process interleaved sequences of visual and textual tokens seamlessly. The result is a system capable of generating summaries that capture abstract concepts, narrative flow, and causal relationships.

While many foundational VLMs were benchmarked on general tasks like visual question answering (Min et al., 2024) or dialogue (Song et al., 2024), their reasoning capabilities are now being explicitly adapted for summarization. For example, models like Video-ChatGPT (Maaz et al., 2024) can be prompted to generate narrative summaries directly from video. Other approaches use LLMs in a multi-step pipeline, such as translating video into intermediate captions and then using a second LLM pass to assess importance and synthesize a summary (LLMVS (Lee et al., 2025)).Frameworks like V2XumLLaMA (Hua et al., 2025) take this further by unifying multiple summarization subtasks under a single, instruction-tuned model. Beyond generation, LLMs are also employed for data augmentation, creating pseudo-ground-truth annotations to train smaller, more specialized models (Islam et al., 2024; Argaw et al., 2024).

This LLM-driven paradigm marks the culmination of the field's convergence. The core task is no longer modality-specific fusion but rather prompting a general-purpose reasoning engine to distill information from a multimodal source. Over time, MMS and VS have evolved from separate, modality-specific pipelines into integrated, attention-driven, and LLM-powered frameworks.

5 Discussion

5.1 Open Challenges

Despite the rapid convergence of MMS and VS, several fundamental challenges must be addressed to create unified, general-purpose systems.

Data Scarcity and Bias Current datasets, as detailed in Section 3, are often confined to narrow domains like news or instructional videos and exhibit significant cultural and linguistic biases (Yuan and Zhang, 2024). This scarcity is especially acute for multilingual and low-resource languages (Lin et al., 2025), which severely limits model generalization. Furthermore, the high cost, intensive labor, and inherent subjectivity of creating large-scale, high-quality multimodal datasets impede the development of robust and comprehensive benchmarks.

Inadequate Evaluation Evaluating generated summaries remains a major hurdle. Existing metrics struggle to holistically assess textual quality, visual salience, and cross-modal consistency (Section 2.2). While recent methods like CLIPScore (Hessel et al., 2021) and mLLM-EVAL (Zhuang et al., 2024) represent progress, they are often too coarse-grained to detect subtle but critical factual errors or misalignments between modalities. Consequently, human evaluation remains the gold standard, but its high cost and low scalability make it impractical for large-scale or long-form content.

Long-Context Modeling Multimodal summarization often demands reasoning over hours of video or documents comprising thousands of tokens (Chandrasegaran et al., 2024). The capacity of current models to maintain context, track narrative arcs, and identify key moments over such long durations remains largely unproven, as they are typically benchmarked on short-form content. Even with advances in long-context architectures, models face a difficult trade-off between computational efficiency and contextual completeness, often leading to the omission of salient information.

Cross-Modal Reasoning and Faithfulness Ensuring that a summary is semantically faithful to all source modalities is a core challenge. Models are prone to hallucinating content unsupported by visual evidence or misinterpreting images when textual context is ambiguous (Wan and Bansal, 2022). Achieving factual consistency requires fine-grained alignment and robust visual-semantic reasoning capabilities that current models only approximate (Li et al., 2018). This issue is particularly critical in high-stakes domains like healthcare, law, and education, where faithful summarization is not just desirable but essential (Zhang et al., 2022b).

5.2 Future Directions

Abstractive Multimodal Summarization While abstractive text summarization is well studied, extending this to non-textual modalities remains a largely unexplored challenge. True abstractive generation—creating novel visual narratives, coherent infographics, or synthesized video clips from source content—is a largely unexplored domain. Progress will depend heavily on foundational advances in generative AI and cross-modal synthesis (Xing et al., 2024).

Novel Evaluation Paradigms Developing multimodal evaluation metrics that jointly assess content selection, factual alignment, and cross-modal grounding is critical. Leveraging multimodal LLM-as-judge (Zhuang et al., 2024) with explicit reasoning steps, combined with task-specific benchmarks for factuality and coherence, may help bridge the gap between automated scoring and human judgment. Human-AI collaborative evaluation that combines automatic tools with expert judgment for more reliable assessment and attribution methods that enables source-to-summary traceability across modalities are also promising directions.

Hierarchical Video Modeling To tackle long-form content, hierarchical and graph architectures that process information at multiple granularities—from local segments to global narratives—are essential (Zhang et al., 2022a; Yuan et al., 2025). This approach, combined with advances in memory-efficient attention, retrieval-augmented generation (RAG), and streaming processing, could enable robust, real-time summarization of lengthy and dynamic content.

User-Centric and Explainable Systems To maximize utility and trust, future systems must be more interactive and transparent. This involves building controllable systems where users can specify summary length, style, focus, or modality balance, likely achieved via instruction tuning or reinforcement learning. It also includes developing explainable systems that provide justifications for their output and trace information back to the source, fostering user trust and enabling easier debugging. Integrating query-based personalization (Li et al., 2023a) will further enhance the practicality of these systems in real-world applications.

6 Conclusion

This survey has bridged the traditionally distinct yet increasingly convergent fields of multimodal and video summarization, offering a unified analysis of their common tasks, datasets, evaluation metrics, modeling approaches, and future trajectories. We highlight the pivotal role of Vision-Language Models in accelerating this convergence. Continued research within this integrated framework is paramount for developing intelligent systems capable of holistically understanding and summarizing the rich tapestry of multimodal information.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: A visual language model for few-shot learning. In *Adv. Neural Inf. Process. Syst.*, volume 35, pages 23716–23736.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021a. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2021b. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE international symposium on multimedia (ISM)*, pages 226–234. IEEE.

Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung. 2024. Scaling up video summarization pretraining with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8332–8341.

Kumbhar Atharva, Kulkarni Harsh, Mali Atmaja, Sonawane Sheetal, and Mulay Prathamesh. 2023. The current landscape of multimodal summarization. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 797–806.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proc. Assoc. Comput. Linguist. Workshop*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst.*, pages 1877–1901.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A. Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. Wikiweb2m: A page-level multimodal wikipedia dataset. *Preprint*, arXiv:2305.05432.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proc. 49th Annu. Meet. Assoc. Comput. Linguist.*, pages 190–200.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4046–4056. Association for Computational Linguistics.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.*, 32(1):56–68.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conf. North Am. Chap. Assoc. Comput. Linguist.*, pages 4171–4186.

Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2019. Summarizing videos with attention. In *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, pages 39–54. Springer.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and adversarial learning for video summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1579–1587.

Zheng Fu, Yubo Wang, Xudong Yang, Chong-Wah Xu, and Yuncheng Zhao. 2021. Mm-avs: A large-scale multi-modal audio-visual scene-aware dataset for weakly-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1769.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. Supert: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. *arXiv preprint arXiv:2005.03724*.

Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. 2020. Classification of important segments in educational videos using multimodal features.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

B. Gong, W.-L. Chao, K. Grauman, and F. Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2069–2077.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Superv. Seq. Labell. with Recur. Neural Netw.*, pages 37–45.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 505–520. Springer.

Mingfei Han, Xiaojun Chang, Heng Wang, and Linjie Yang. 2023. Shot2story20k: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*.

Michael Hanna and Ondrej Bojar. 2021. A fine-grained analysis of bertscore. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 507–517. Association for Computational Linguistics.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proc. Conf. Empir. Methods Nat. Lang. Process.*, pages 7514–7528.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2025. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3599–3607.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

Zhong Ji, Fang Jiao, Yanwei Pang, and Ling Shao. 2020. Deep attentive and semantic preserving video summarization. *Neurocomputing*, pages 200–207.

Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(6):1709–1717.

Hao Jiang and Yadong Mu. 2022. Joint video summarization and moment localization by cross-task sample transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16388–16398.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. 2013. Large-scale video summarization using web-image priors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2698–2705.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

Mateusz Krubiński and Pavel Pecina. 2023. Mlask: Multimodal summarization of video-based news articles. In *Findings of the association for computational linguistics: EACL 2023*, pages 910–924.

Shamit Lal, Shivam Duggal, and Indu Sreedevi. 2019. Online video summarization: Predicting future to better summarize present. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 471–480.

Min Jung Lee, Dayoung Gong, and Minsu Cho. 2025. Video summarization with large language models. *arXiv preprint arXiv:2504.11199*.

Y. J. Lee, J. Ghosh, and K. Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 1346–1353.

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, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. 2023a. Progressive video summarization via multimodal self-supervised learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5584–5593.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4152–4158. ijcai.org.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020c. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

Yake Li, Junnan Liu, Jie Lu, and Yue Xu. 2020d. Vmsmo: Learning video and music story matching with cross-modal embedding. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, pages 1989–1997.

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

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 26:5548–5560.

Kaiying Kevin Lin, Hsiyu Chen, and Haopeng Zhang. 2025. Formosanbench: Benchmarking low-resource austronesian languages in the era of large language models. *arXiv preprint arXiv:2506.21563*.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. What is that talk about? a video-to-text summarization dataset for scientific presentations. *arXiv preprint arXiv:2502.08279*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Tengchao Lv, Lei Cui, Momcilo Vasilijevic, and Furu Wei. 2021. Vt-ssum: A benchmark dataset for video transcript segmentation and summarization. *arXiv preprint arXiv:2106.05606*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of TV shows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8272–8291, Bangkok, Thailand. Association for Computational Linguistics.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245.

Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. Clip-it! language-guided video summarization. In *Adv. Neural Inf. Process. Syst.*, volume 34, pages 13988–14000.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the evaluation of video summaries. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7596–7604.

Pinelopi Papalampidi and Mirella Lapata. 2023. Hierarchical3D adapters for long video-to-text summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320, Dubrovnik, Croatia. Association for Computational Linguistics.

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. Refinesumm: Self-refining mllm for generating a multimodal summarization dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13773–13786.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, and 1 others. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aishwarya Ramakrishnan and Chun-Kit Ngan. 2022. A hybrid video-to-text summarization framework and algorithm on cascading advanced extractive- and abstractive-based approaches for supporting viewers' video navigation and understanding. In *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 36–39.

Andrea Sanabria, Jiri Matas, Aki Malm, and Edward Grefenstette. 2019. How2: A large-scale dataset for multimodal language understanding, including sports content. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 423–431.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (EMNLP)*, pages 9–14.

Yassir Saquil, Da Chen, Yuan He, Chuan Li, and Yong-Liang Yang. 2021. Multiple pairwise ranking networks for personalized video summarization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1718–1727.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Zirui Shang, Yubo Zhu, Hongxi Li, Shuo Yang, and Xinxiao Wu. 2025. Video summarization using denoising diffusion probabilistic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6776–6784.

Jaewon Son, Jaehun Park, and Kwangsu Kim. 2024. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18847–18856.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Xuemeng Song, Liqiang Jing, Dengtian Lin, Zhongzhou Zhao, Haiqing Chen, and Liqiang Nie. 2022. V2P: vision-to-prompt based multi-modal product summary generation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 992–1001. ACM.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 5179–5187.

Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. Tldw: Extreme multimodal summarization of news videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(3):1469–1480.

Yunlong Tang, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. 2022. Multi-modal segment assemblage network for ad video editing with importance-coherence reward. In *Proceedings of the Asian Conference on Computer Vision*, pages 3519–3535.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the blanc: Human-free quality estimation of document summaries. *arXiv preprint arXiv:2002.09836*.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *EMNLP 2022*.

Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. 2019. Stacked memory network for video summarization. In *Proceedings of the 27th ACM international conference on multimedia*, pages 836–844.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. 2024. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

S. Yeung, A. Fathi, and L. Fei-fei. 2014. VideoSET : Video summary evaluation through text. `https://arxiv.org/abs/1406.5824`. ArXiv preprint arXiv:1406.5824v1.

Dian Yu, Chen Xu, Yue Zhang, and Zhiyang Jiang. 2021. Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2232–2242.

Haohan Yuan, Sukhwa Hong, and Haopeng Zhang. 2025. Strucsum: Graph-structured reasoning for long document extractive summarization with llms. *arXiv preprint arXiv:2505.22950*.

Haohan Yuan and Haopeng Zhang. 2024. Domainsum: A hierarchical benchmark for fine-grained domain shift in abstractive text summarization. *arXiv preprint arXiv:2410.15687*.

Yuan Yuan, Haopeng Li, and Qi Wang. 2019. Spatiotemporal modeling for video summarization using convolutional recurrent neural network. *IEEE Access*, pages 64676–64685.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2022a. Hegel: Hypergraph transformer for long document summarization. *arXiv preprint arXiv:2210.04126*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022b. Improving the faithfulness of abstractive summarization via entity coverage control. *arXiv preprint arXiv:2207.02263*.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2025. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, 57(11):1–41.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *Proc. Eur. Conf. Comput. Vis.*, pages 766–782. Springer.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022c. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11676–11684. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proc. Int. Conf. Learn. Represent.*

Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

Yujia Zhang, Michael Kampffmeyer, Xiaoguang Zhao, and Min Tan. 2019. Dtr-gan: Dilated temporal relational adversarial network for generic video summarization. In *Proc. ACM Turing Celebr. Conf. - China*, pages 1–6.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022d. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2017. Hierarchical recurrent neural network for video summarization. In *Proc. ACM Int. Conf. Multimedia*, pages 863–871.

Bin Zhao, Xuelong Li, and Xiaoqiang Lu. 2018. Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7405–7414.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018a. MSMO: multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4154–4164. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

Qi Zhu, Tao Li, Xiangyu Zhang, Wei Lu, Hsin-Yu Wang, and Baotian Hu. 2018b. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4154–4164.

Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. 2020b. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962.

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. 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 7768–7790.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.